Summer 9-1-2014

Anger and Punishment: Natural History and Normative Significance

Isaac Thane Wiegman
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Philosophy
Philosophy-Neuroscience-Psychology Program

Dissertation Examination Committee:
Ron Mallon, Chair
John Doris, Co-Chair
Carl Craver
Julia Driver
Tammy English
Daniel Kelly

Anger and Punishment: Natural History and Normative Significance

by

Isaac Wiegman

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2014

St. Louis, Missouri

Table of Contents

Acknowledgements

I have a great deal of gratitude for many people without whom this dissertation could not have been written. I mention only a few of them here.

First, my wife Juli has been a steady source of support and care through the entire process, pulling me through the inevitable low grade depression and mood swings of technical writing as only she could. Moreover, she has engaged this work in a way that only an insightful counselor could, with loads of practical knowledge about emotion. Other sources of familial support came from my mom and dad, brother and sisters. In particular, I have had a lot of fun discussions with Caroline Wiegman about this work, and Valerie Wiegman drew the illustration on page 112. This is not to mention a large circle of St. Louis friends who have been like family to Juli and I.

Second, my advisors Ron Mallon and John Doris knew what to do with me at various points when I didn't really know what to do with myself. They kept me honest, provided reams of helpful feedback, provided hundreds of hours of face to face guidance and patiently waited for me to produce the best work that I could.

Third, thanks to members of my committee. At the beginning of this project, I had a number of helpful discussions with Dan Kelly, whose book inspired this project. Tammy English agreed to participate at the end on short notice, and she helped me to see some unnecessary liabilities in the foundations of the project. Julia Driver provided very helpful feedback on chapters 1 and 4. Carl Craver has had a huge influence on me through his courses, his feedback and through our intermittent conversations.

Fourth, several other members of the Philosophy Department provided feedback on previous versions of each chapter. Without Anya Plutynski's feedback, chapter 3 would be much more confusing. The same goes for Lizzie Schecter's feedback and encouragement on chapter 4 and Anne Margaret Baxley's feedback and encouragement on chapter 1. I had loads of good discussion with Ian Tully, Felipe Romero, Taylor

Murphy and Tom Wysocki on chapters 1, 2, 3 and 4 respectively, and all provided helpful feedback on the respective chapters.

Fifth, thanks to the Philosophy-Neuroscience-Psychology Dissertation Preparation Seminar, the members of which provided feedback on several different chapters at different stages of disorder, thanks especially to Mike Dacey, Nazim Keven, and Lauren Olin.

Last but not least, thanks to Nate Adams, without whom I quite literally would not have made it through graduate school. He has been a constant source of clarity, on matters both philosophical and personal. Together, we shared meals, struggles, doubts, joys, ideas, and sci-fi readings. I wouldn't mind a zombie apocalypse so long as he was part of my tribe.

-To Juli, Sofia and Anabel

I believe very strongly…that no one could ever deserve to suffer. Of the people in whose moral judgment I have the most confidence, some disagree. When some wrong-doers suffer, these people believe, this suffering is in itself good, or at least not in itself bad. Though this belief seems to me mistaken, I would be greatly relieved if I could explain why these people are making this mistake. This may be one of the cases in which an evolutionary explanation helps to undermine what it explains. This retributive belief may seem to justify certain natural reactive attitudes, such as an angry desire to hurt or the withdrawal of good will. These attitudes are like some simpler emotions that are had by the animals that are most like us. If evolution can explain why many people have these reactive attitudes, that might give some support to the view that these attitudes, and the widely held belief that such attitudes are justified, are not responses to reasons.

-Derek Parfit (2011)

Introduction: Circumscribing the Phenomena

Punishment, whether institutional (e.g. by governments) or interpersonal (e.g. by parents) usually involves the imposition of hard treatment on the punishee. These hard treatments can include the curtailment of certain liberties, such as freedom of movement and association, as well as privacy. Prima facie, people have rights to these liberties. Moreover, hard treatments expend resources and cause suffering. Given its costs and its curtailment of liberties, it is no wonder that philosophers and legal theorists have spilled a great deal of ink developing justifications for punishment.

Two *positive* considerations are usually given as justification for punishment (as opposed to *negative* considerations that constrain or limit punishment). Consequentialist considerations are forward-looking or *prospective* in nature, appealing to the good outcomes that will result from punishing a transgressor, outcomes like general deterrence, rehabilitation, or incapacitation. The other kind of consideration is *retributive*, appealing to what a transgressor deserves given their transgression.[1] These retributive considerations have weight only if the punishment of transgressors has some positive value aside from its consequences.[2]

Retributive considerations for punishment have considerable intuitive appeal (as I demonstrate below), and they factor strongly into justifications (popular or otherwise)

[1] There are other considerations relevant to punishment that do not fall under either of these headings. For instance, one might have a right to make threats against those who might violate one's rights, and the right to punish might follow from this right to threaten (Quinn 1985). This kind of justification is not of interest to me here. Note however, that to the punishee, even this justification will seem retributive in this sense, it "…is a deprivation inflicted on someone found guilty [of violating a right], and not on anyone else, and it is imposed solely because of that finding." (Bedau and Kelly 2003) In this sense, the procedural justification of punishment warranted by this theoretical justification comports with the retributive motive that I characterize below.

[2] This positive retributive consideration is often contrasted with the *negative* retributive consideration (also contrary to straightforward consequentialist evaluations of action) that the innocent should not be punished. I mention this possibility only to point out that negative considerations are not my focus here. Rather, my focus is on the motivation to punish the deserving (e.g. anger), which I suspect to be distinct from the motivation to avoid punishing the guilty (e.g. anticipated guilt or regret). Henceforth, I will use "retributive considerations" and "retributive reasons" to refer to only to *positive* retributive considerations in favor of punishment.

for harsh treatments such as incarceration and capital punishment (see e.g. Moore 2010; Pojman 2005). Thus, there is good reason to evaluate this kind of justification for punishment, to see what role, if any it ought to play in reasoning about punishment. One way of supporting retributive considerations is to appeal to widespread retributive intuitions. One purpose of this dissertation is to undermine these intuitions as evidence in support of retributive considerations.

Doing this requires more clarity about retributive considerations. Insofar as retributive considerations can lead agents to choose punishment over a range of alternatives, it makes sense to think of retribution as a motive for action. In this introductory chapter, I characterize the structure of this motive. Thus, in the first section I review some of the philosophical issues surrounding retribution, both clarifying the relevant concept of retribution and pointing out the philosophical import of the question on which I focus: why are humans motivated to punish even apart from the good consequences that punishment can bring about? In the second section, I review some of the psychological work surrounding punishment, highlighting its connection with anger and outrage and pointing out a plausible role for anger in motivating retributive punishment. In the third section, I say more about what anger is by describing some well-established components of the anger syndrome and describing the most prominent cluster of strategies for explaining these phenomena, *basic emotion theory*. In the final section, I summarize the chapters to come and their role in explaining and evaluating the retributive motive.

1. From retributive considerations to retributive motives

First, retributive considerations favor punishment, which immediately raises the question of what punishment is. Here, I am concerned specifically with moral

punishment, which has three important features (cf. Walen 2014) for my purposes.[3] First, punishment involves the imposition of a cost, usually in the form of hard treatment. Second, punishment is intentional. If I impose a cost on someone by accident or as an unintended (but perhaps foreseen) side effect of some other action, that cost is not easily understood as a punishment. Third, punishment is a response to wrongdoing. If someone imposes a cost on someone even though one believes him or her to be innocent of wrongdoing, such action does not fit well with other instances of punishment.[4] [5]

So then, how do consequentialist and retributive considerations justify the imposition of hard treatment in response to wrongdoing? On one understanding, a consequentialist evaluation of action (as right or wrong) "...depends only on [the action's] consequences (as opposed to the circumstances or the intrinsic nature of the act or anything that happens before the act)." (Sinnott-Armstrong 2003) In other words, a consequentialist justification for punishment will refer to the ability (or likelihood) of punishment to bring about good or valuable outcomes. The good outcomes that justify punishment might vary (e.g. restraint or rehabilitation of the offender, deterrence of crime) along with the set of values in terms of which outcomes are evaluated (e.g. pleasure, pain, satisifed preferences, or even virtue and vice). Insofar as punishment brings about a better outcome than alternatives (as determined by the amount of value assigned to the outcome), there is a reason to punish. From a psychological perspective,

[3] These criteria have the flavor of a definition. Nevertheless, I do not think of these criteria as a definition or as necessary and sufficient conditions for an act to answer to the concept of punishment. Rather, these criteria are helpful in roughly circumscribing the phenomenon of punishment, which is properly understood as a unified phenomenon because all of its instances share common explanatory elements.

[4] Many legal philosophers follow Feinberg (1970) in claiming that punishment is distinct from other penalties (e.g. fees for parking violations) in that it has the function of condemning the wrongs that it penalizes. For my purposes, the third criterion, that punishment is a response to wrongdoing, distinguishes punishment from other penalties, so long as wrongdoing is understood in terms of *mala in se* offenses (offenses that are prohibited because they are morally wrong or objectively offensive) rather than *mala prohibita* offenses (mere offenses against what the law prohibits for other reasons).

[5] In chapter 4, I give a behavioral definition of punishment for the purposes of identifying a specific strategy for social interaction, and I argue that retributive motives are an adaptation for bringing about that behavioral strategy. While I believe there is an important connection between moral punishment and punishment as a behavioral strategy, it is important to keep these notions distinct.

people are moved by consequentialist motives when their actions are instrumental for bringing about good outcomes.

The forward-looking focus of consequentialism (on future outcomes) contrasts with the backward-looking focus of retributivism: giving the offender what she *deserves* given the nature of the past offense. This contrast captures a central bone of contention that divides two traditional ethical theories: the relevance of desert for evaluating states of affairs. A central tenet of the Kantian tradition is that each person should have happiness in accordance with their virtue, presumably because happiness is what the virtuous deserve. By contrast, the purest kind of (hedonistic) utilitarianism (the forerunner of contemporary consequentialist theories) holds that an action is morally justified only if it leads to the best outcome, defined by the sum total of happiness for everyone and *even if happiness is distributed broadly among the vicious or scarcely among the virtuous*.

Retributive justifications for punishment are squarely within the Kantian tradition. They focusing on which punishment is appropriate given the wrongdoing that one committed. Moreover, this sense of appropriateness is usually understood in terms of desert. In one sense, this motive to punish is deeply mysterious: "[it] appears to be a mysterious piece of moral alchemy in which the combination of the two evils of moral wickedness and suffering are transmuted into good." (Hart 1967, 234–235) Part of the mystery here lies in the assumption that suffering is what retributive punishment aims at or that suffering is the object of desert (what a person deserves for wrongdoing). Nevertheless, even if one assumes that what an offender deserves for wrongdoing is *hard treatment* (regardless of whether this leads to suffering) there remain intractable questions about why a person's wrongdoing makes it appropriate to impose hard treatment.

Several answers to this question have been attempted. Some are based on rectifying the advantages that people gain from acting wrongly, advantages that are not deserved (Morris 1968). Others are based on communicating censure or moral condemnation for wrongdoing, whether or not this communication has a good outcome (e.g. Duff 2001). Still others focus on vindicating the moral status of victims, which is compromised by wrongdoing (Hampton 1992). This vindication is thought of as an appropriate response to wrongdoing, independently of whether it brings about any future benefits.

While all of the justifications above are retributive in nature, there are also consequentialist justifications for giving offenders what they deserve. For instance, if we want to construct a practice of punishment that deters would-be transgressors, the best way to do this might be to make would-be criminals believe that we will punish them in reaction to their culpable wrongdoing and *not only* when the punishment would have favorable consequences. This way, they will not be able to bargain their way out of punishment by doing something that would make it more profitable or beneficial for us not to punish them.[6] Nor would they be able to choose a fortuitous transgression for which the deterrent benefit would exceed the cost of prosecution and punishment. It is difficult to think of a better way to deter premeditated wrongdoing than to make a public commitment to punish primarily based on culpable wrongdoing rather than on the consequences punishment will actually have in a given instance.

Nevertheless, as a psychological phenomenon, retributive motives do not depend on any of these justifications. The motives themselves preexist any attempts by philosophers justify them or to make them intelligible.[7] Moreover, there remains a strong feeling that wrongdoers deserve to be punished whether or not one accepts any of

[6] As an example, Jason Beckman offered $19 million for leniency before being sentenced for a Ponzi scheme: http://www.startribune.com/local/183100491.html

[7] For an interesting historical explanation of how these motives became enshrined in western law, see the final chapter of Daly and Wilson (1988).

these justifications. Let us accept for a moment the consequentialist justification for a retributive practice of punishment (as expressed in the previous paragraph). We can imagine a case in which a murderer is publicly convicted of his crimes, and given the death penalty (let us suppose that this is the best deterrent for murder). Nevertheless, in this case, a merciful government official fakes the execution. Moved by the official's mercy, the murderer finds herself completely morally reformed and becomes incapable of committing another murder. She assumes another identity and lives a prosperous life. So long as the truth were secure from detection, all the good outcomes of punishment would have been achieved in this case (e.g. general deterrence, rehabilitation, prevention of future crimes by the offender). Nevertheless, many will find it galling to imagine the murderer getting away with her crime.

No matter how compelling one finds consequentialist justifications of punishment, and no matter how confident one is that the good outcomes of punishment are secured by other means, there is still an obstinate feeling that it is wrong to let the guilty go unpunished. I believe that the same feeling will tend to persist regardless of how compelling one finds other prominent attempts to justify it or to make it intelligible. For my purposes, the point is twofold. There is an interesting phenomenon to explain and a psychological motive to evaluate even aside from the patterns of reasoning that philosophers use to *support* retributive reasons to punish.

From a psychological perspective, the most straightforward way to characterize the motive is not the patterns of reasoning that support it, but rather the patterns of action for which it is responsible. If one is moved by retributive motives, then there are cases in which one would punish, or report that punishment is fitting, even though one knows that punishment would not bring about any future benefit. Moreover, retributive motives explain this pattern, because they place value on punishment as appropriate, fitting, or deserved (however these notions happen to be fleshed out or rationally

supported) in response to *past* wrongdoing (rather than because of its *future* consequences).

Importantly, the distinction I have drawn between retributive and consequentialist considerations is not captured by a key difference between consequentialist moral theories and duty-based, or deontological, theories. For some philosophers (R. Nozick 1974; Scheffler 1994), what distinguishes deontological theories is the acceptance of constraints on action that limit the permissibility of bringing about good outcomes or the obligation to do so. To accept such constraints is to deny one central consequentialist commitment: that it is morally right (perhaps even required) to bring about the better outcome. On my usage, retributivist considerations need not be subject to such constraints. For instance, someone can accept that the better outcome is the one in which the greater number of guilty offenders is punished and choose to bring about that outcome *even if it would require withholding deserved punishment from a smaller number of equally guilty offenders*.[8] More concretely, I could accept that there are retributive reasons to punish Maria for her crime, but nonetheless forego punishing Maria in order to punish Sally and Rebecca for their crimes (also for retributive reasons). In other words, accepting retributive reasons for punishment does not entail that one is obligated to punish the guilty whenever it is in one's power to do so or independently of the consequences of *not* punishing (for an alternative view, see Berman 2011).[9]

[8] Essentially, I am saying that a thoroughgoing retributivist need find nothing wrong with the practice of pleabargaining.

[9] As a further clarification, retributive considerations may sometimes violate the central tenet of consequentialism (that one is permitted or required to bring about the better outcome), but they need not do so as a matter of definition. For instance, by punishing a transgressor as they deserve I might diminish the amount of happiness in the world by making the transgressor less happy and without making anyone else any more happy. Now suppose I believe that I ought to try and promote happiness, but also that retribution is required of me in this case. I can consistently hold both of these beliefs if I also accept that the pursuit of retribution constrains the permissibility of bringing about the better outcome (e.g. in which the transgressor is more happy rather than less). By contrast, I might adopt a more pluralistic theory of value according to which retributive justice is itself a value worth promoting. On this kind of view, a world in which the transgressor is punished could be a better world than one in which she is not, even after factoring in her diminished happiness. In that case, to enact retribution is consistent with believing that it is always

In sum, the psychological entity that I am interested in explaining and evaluating is the retributive motive. It is a motive to impose hard treatment in response to wrongdoing. It is best captured by the pattern of actions it produces: punishment (or punishment of a certain kind or severity) when it will obtain no future benefit (or less future benefit than an alternative punishment). It produces those patterns of action because it causes people to find punishment fitting (in some sense) as a response to past wrongdoing (where fittingness is understood independently of the consequences that punishment brings about). Like duty-based ethical theories in the tradition of Kant, the focus of retributive reasons on the past is often understood in terms of desert. Unlike some contemporary expressions of duty based theories, it does not, by definition, include any constraint on bringing about good consequences.

2. The Phenomena of Punishment and Retribution

Fortunately, there is a wealth of studies in psychology and in behavioral economics relevant to retributive motives for punishment. Here I review some of this work. Importantly, this work suggests that anger influences many peoples' judgments and actions concerning punishment, and may be responsible for the retributive bent of those judgments.[10]

---

permissible to bring about the better outcome. From both perspectives, the act of punishment is naturally understood as retribution, yet only on one of these perspectives does the justifiability of retribution require setting limits on my pursuit of the best outcome. While in the latter case the decision to punish might seem to have a purely consequentialist justification, the judgment actually has a non-consequentialist dimension. The world may be a better place because I punished the transgressor, but it is not a better place *only because of the consequences of my action*. This is because on my view "consequence" is understood in isolation from what happened before the act of punishment (cf. Sinnott-Armstrong above). So long as consequences are understood in this way, the difference between a world in which the transgressor is punished and one in which she is not cannot *only* be a difference in the consequences of my action, but also in their relation to what came before my action (the transgression). Moreover, one cannot understand the stated justification for punishment in isolation from what came before the act. Thus, one cannot understand this justification solely in terms of consequentialist reasons for punishment.

[10] There are other emotions aside from anger that might influence the development of retributive intuitions. For instance, when we consider the perspective of a transgressor, we are sometimes overcome with the guilt that he or she must feel. We can easily imagine that some transgressors would feel guilt or remorse even if they had not been caught. It will not always assuage a transgressor's guilt to convince herself that the overall outcome will be better if she is not punished. This is just to say that sometimes, transgressors are in agreement with their victims that they deserve to suffer and that their suffering would be good in and of

One line of evidence concerns the severity of punishment that people assign to offenders when considering various scenarios. This research licenses some inferences about which aspects of scenarios influence the severity of punishment assigned. There two key predictions that one can make about peoples' punishment decisions based on the contribution of consequentialist and retributive motives to punish. First, if punishment decisions are influenced by retributive motives, then the severity of punishment will change in response to factors that influence desert (e.g. culpability, seriousness of crime). Second, if punishment decisions are influenced by consequentialist motives, then the severity of punishment will change in response to changes in the consequences of punishment (e.g. its deterrent value and its foreseeable side effects). For instance, probability of detection determines the severity of punishment required to achieve a certain level of deterrence. So if someone's punishment judgments are motivated by the aim of deterrence, then the severity of punishment should increase as the probability of detection decreases.

This prediction has not born out. Over the course of nine studies, Baron and Ritov (2009) asked both judges and internet participants to assign penalties for various crimes and also had them rate the seriousness of the crime and their anger at the crime (in some of the experiments). In almost every case, the seriousness of the crime or anger in response to it were much better predictors of the severity of punishment than the probability of detection for the crime. Only a few participants' severity assignments tracked the probability of detection. Moreover, only when probability of detection was highly salient did it influence severity of punishment in the direction predicted by consequentialist motives. Contrary to their predictions, Baron and Ritov found that even when participants took on a policy-making perspective as opposed to making judgments

---

itself. This belief is a product of their guilt, but I suspect that the evolutionary history of guilt has been shaped by its relationship to anger, such that the natural history of anger will end up explaining why guilt also lends itself to this intuition (see DeScioli and Kurzban 2009).

about a specific violation (e.g. “This item is about how future offenses should be penalized” as opposed to “about an offense already committed.”, p. 572), lower probability of detection did not lead to increased severity of punishment.

Carlsmith (2008) explored the influence on punishment of a wider range of consequentialist considerations (e.g. “the publicity of the crime and subsequent punishment, the frequency of the crime, the likelihood of similar crimes in the future, the likelihood of detecting the crime, and the likelihood of catching the perpetrator.” Carlsmith, 2008, p.124) and retributive considerations (e.g. “the severity of the harm, the moral offensiveness of the behavior, the intent behind the action, the blameworthiness of the offender, and whether or not the offender was acting in a responsible manner.” Carlsmith, 2008, p.123-124). As a partial replication of previous work with colleagues (Darley, Carlsmith, and Robinson 2000; K. M. Carlsmith, Darley, and Robinson 2002b), he found that consequentialist considerations were poor predictors of peoples’ actual sentencing decisions, whereas retributive considerations were strong predictors. Even though peoples’ stated motives for punishment usually focused on deterrence, these reports did not correlate at all with peoples’ actual decisions.

In another study, Carlsmith (2006) investigated peoples’ decision making process by allowing participants to selectively and sequentially access different information about a given criminal offense, probing for severity of sentence and confidence in sentence after each selection. Some of the information concerned retributive considerations (i.e. “Magnitude of harm”, “Perpetrator intent”, “Extenuating circumstances”). Other information had more to do with consequentialist considerations (i.e. “Likelihood of violence”, “Prior record”, “Self-control”, “General frequency”, “Detection rate”, “Publicity”). The key result was that people frequently chose to access information relevant to retribution prior to accessing information relevant to incapacitation or general deterrence. Moreover, Carlsmith reported that “…retribution

information improved confidence more than did incapacitation information." (K. M. Carlsmith 2006, 446)

Some of these studies (including Baron and Ritov 2009) indicate that anger is connected to retributive judgments. For instance, Carlsmith et al (2002a) found that in response to vignettes about punishment, moral outrage ratings were a strong predictor of punishment and mediated the influence of retributive considerations (i.e. absence of mitigating factors and seriousness of offense) on those judgments.

Importantly, none of these studies manipulated anger to measure its effects on punishment judgments. However, others (J. S Lerner, Goldberg, and Tetlock 1998; Goldberg, Lerner, and Tetlock 1999) have done just that. In one of these experiments (J. S Lerner, Goldberg, and Tetlock 1998), one set of participants, the anger induction group, watch a video depicting a bully and an accomplice who assault and humiliate a teenager. Another set of participants, the control group, watched a video of abstract colors and shapes with negligible emotional content. Afterwards, participants read vignettes describing hypothetical harms to the participant and rated the degree to which the perpetrators ought to be punished. The punishment ratings of the anger induction group were higher than controls, demonstrating that incidental anger leaks into judgments about punishment.

A more direct source of evidence about the influence of (non-incidental) anger comes from experiments on the ultimatum game (UG). In this game, one player, the proposer, receives a sum of money and is instructed to choose how much of it to share with another player, the receiver. The receiver then has the option of accepting or rejecting the offer. If she accepts, then both players get the portion of money assigned by the proposer. If she rejects, then neither player gets any money. Prima facie, peoples' performance on one-shot UGs looks retributive. Low offers (e.g. $2 out of $10) are frequently rejected, even when people are playing with real money. Thus, receivers

frequently impose costs on proposers, even though (due to fact that the interaction is not repeated) there *appears* to be no visible benefit. This is just what we might predict if receivers have a retributive motive for rejecting low offers.

Of course, appearances can be misleading. From this crude behavioral description of the rejection, we cannot infer much about the receiver's motivation for rejecting the offer. Perhaps receivers want to enforce fairness norms that might benefit many people in the long run. Likewise, this behavioral description of the receiver's rejection does not yet fit with the description of punishment above, because it is unclear whether the receiver interprets the proposer's behavior as wrongdoing.

Fortunately, there is an abundance of data on the UG to resolve these ambiguities. Receivers do judge low offers to be unfair (e.g. Pillutla and Murnighan 1996), and as such, this is good reason to suppose that they believe the proposer acted wrongly (by acting unfairly). While their unfairness ratings tend to correlate strongly with rejections, unfairness ratings do not entirely explain rejections. For instance, rejections of unfair offers are correlated with activation of the anterior insula (as measured by changes in BOLD signal detected via fMRI), an area of the brain associated with anger and disgust (Sanfey et al. 2003). Additionally, UG participants more frequently reject low offers if anger is induced (e.g. by journaling about an angering event in their past) prior to playing the UG (Srivastava and Espinoza 2009). In some of these anger induction experiments, participants are alerted to their emotional state and instructed to make sure that induced anger does not influence their performance on the UG. In these experiments, the rejection of unfair offers goes down even though their judgments of

unfairness remain relatively constant. So anger seems to motivate rejections of low offers over and above the mere judgment that low offers are unfair.[11]

Of course, these results still only deal with incidental anger. In another study (Fabiansson and Denson 2012), participants were angered by a confederate who criticized a speech that they had just given. These participants then participated in several one-shot, computerized UGs with the speech counterpart and two other players (the offers in the game were actually automated, and a photograph of the "other player" was shown during play). Participants were more likely to reject unfair offers from the person who criticized their speech than from the two other fictitious players. The difference in rejection rate is not easily explained without referring to anger directed at the speech counterpart, nor is it easily explained by referring to any future outcome that participants hoped to bring about.[12] Rather, the clearest explanation is that participants were angry at the speech counterpart, and this anger made them more likely to reject offers from the speech counterpart, regardless of whether this would bring about a future benefit.

At this point, two related worries arise. First, one might point out that people also make cold and detached judgments concerning punishment. This raises a question of how anger can influence punishment judgments even when people are not in the immediate grip of anger. I do not want to deny the possibility of cold punishment judgments. Nevertheless, even in the questionnaire studies above in which participants approach vignettes impartially, they still report feelings of (or perhaps dispositions toward) anger or moral outrage at offenses, and it looks like these feeling inform their judgments about punishment. I say more about this below.

[11] Moreover, it is not just the negative valence of anger that influences rejection. When sadness is induced and participants are instructed to ignore their sadness, their rejection of low offers does not diminish (Srivastava and Espinoza 2009, 485).

[12] It is not inconceivable that participants wanted to deter the speech counterpart from insulting others, but to me it seems much more likely that the motive was retributive.

Second, some will think that, due to the influence of anger, this behavior in the UG is more like revenge than retribution, concepts which many philosophers keep separate. According to Nozick, "Revenge involves a particular emotional tone, pleasure at the suffering of another, while retribution either need involve no emotional tone or involves another one, namely pleasure at justice being done." (Robert Nozick 1981, 367) Nevertheless, as Zaibert (2006) argues in detail, this supposed difference between punishment and revenge is not only sorely under-defended, but also highly implausible. To see the implausibility, one need only imagine the archetypal *Godfather* Don calling in a hit on someone who has just left his office. We can imagine him having exactly the degree of emotional detachment as a judge pronouncing the death sentence on a convicted murderer. By contrast, we can also easily imagine the judge sentencing, or perhaps the executioner punishing, with all the maniacal fury of a McCoy avenging herself on a Hatfield. In this case, one need not think that the judge or the executioner are *really* acting as avengers rather than punishers.

More importantly, it is empirically false that retribution has no emotional tone. The high association between reports of anger or moral outrage at an offense seems to indicate that even when approaching criminal offenses impartially (as in the questionnaire studies above), there is usually some emotional tone or disposition toward the wrongdoer. Moreover, in one experiment, the degree of activation in brain regions associated with emotion (e.g. the amygdala) varied in proportion to the severity of punishment that subjects assign in response to vignettes describing criminal behavior (Buckholtz et al. 2008). Thus some emotional processing may influence punishment judgments even when those judgments are made from an impartial standpoint.

Additionally, reported emotional states are not restricted to the UG or to the questionnaire studies reviewed above. Consider, for instance, a modification of the UG, called the dictator game (DG). In this version of the game, the receiver just passively

receives the offer of the proposer (the "dictator") and has no option to accept or reject the offer. Nevertheless, when a third party is added to this scenario and given the option of deducting points from the dictator at a cost to herself (e.g. Fehr and Fischbacher 2004), we can observe actions that fit more readily into the phenomenon of punishment. This is because costs are imposed in response to an action that is judged unfair *from an impartial standpoint* (which seems to be the paradigmatic case for many philosophers and legal scholars). Even in this case, the third party reports anger in response to the dictator's offer. Moreover, reported anger predicts punishment and mediates the influence of retributive considerations, such as the dictator's culpability for low offers (Nelissen and Zeelenberg 2009). The point is that reports of outrage and anger don't just accompany *judgments* that someone deserves punishment, nor do they only accompany punishment behaviors that resemble *revenge* (as in the UG), but they also accompany behaviors that are imposed from an impartial perspective (and sometimes at a cost to the punisher). Henceforth, I will call these behaviors impersonal punishment as opposed to the more personal form of punishment observed in the UG.[13]

There probably are interesting psychological differences between impersonal cases of punishment and the kind of "hot" punishment that one finds in the UG. Nevertheless, in the remainder of this section, I argue that the difference between these phenomena is not that one of them lacks any influence from anger. Rather the difference resides in the *nature* of the influence that anger exerts over these distinct phenomena.

Consider some affective differences between impersonal punishment and personal punishment (for instance, in the UG). In one experiement, van t' Wout and colleagues (2006) found that when their participants played the UG (as the receiver and ostensibly against other humans) their skin conductance activity (a measure of affective response) was higher for lower offers and was correlated with rejections. However, when

[13] Usually, this kind of punishment is referred to as "third party punishment".

people played the UG with a computer, neither relationship was observed. Similarly, Civai et al had participants play the UG (also as the receiver) both for their own monetary gain (the *myself* condition) and then on behalf of a third party (the *third party* condition). While rejection rates were very similar between the two tasks, rejection of offers was only accompanied by differences in skin conductance in the *myself* condition. No significant difference was found between rejection and acceptance of offers in the *third party* condition, presumably because low offers did not affect the participant's pay out.

Other studies suggest that, regardless of what participants report, anger and outrage are probably not *experienced* in conjunction with impersonal punishment. For instance, psychologist have yet to identify any cases in which outrage is elicited by the fact that someone violated a moral norm, as opposed to the fact that the violation harmed the subject or someone she cares about (Batson et al. 2007; Batson, Chao, and Givens 2009). At the very least, the difference in skin conductance between personal and impersonal punishment suggests that emotional arousal is lower for impersonal punishment. This might tempt someone to think that personal and impersonal punishment are distinct phenomena and that only impersonal punishment is influenced by anger. The latter inference does not cohere well with the data presented above. If anger does not influence impersonal punishment in some way, then it is difficult to account for reports of anger and outrage in conjunction with punishment judgments in questionnaire studies or punishment behaviors in economic games.

Moreover, there are interesting neurological and genetic connections between the phenomena of impersonal punishment and personal punishment that should be accounted for. For instance Strobel and colleagues (Strobel et al. 2011) analyzed brain activation of participants who engaged in both an impersonal punishment task (as a third party in the DG) and in a personal punishment task (similar to the UG). Activation

in the nucleus accumbens and nucleus caudatus (brain regions associated with reward) differed significantly depending on whether the participant punished or not (as did other brain regions associated with emotion, e.g. the amygdala). Moreover, the difference was observed in both the impersonal and personal punishment tasks (though there was a difference in the magnitude of activation between personal and impersonal punishment). This suggests that the motivation to punish may be similar across impersonal and personal punishment.

Strobel and colleagues also found genotype-specific differences in brain activation based on the contrast between punishment and non-punishment (across the impersonal and personal punishment tasks). That is, a specific allele for a gene controlling dopamine turnover predicted the difference in activation in several regions (including the nucleus accumbens and the amygdala) between punishment and non-punishment, regardless of whether punishment was impersonal or personal.

While the data is suggestive, I suspect that it is too soon to say with any confidence what the similarities and differences are between impersonal and personal punishment in economic games (and elsewhere if these results can be generalized). However, I suspect that anger influences the development of impersonal punishment, even if anger is not manifested or experienced during impersonal punishment. Specifically, I suspect that anger influences the development of the *response-dependent* category of *the outrageous*. Response-dependent categories are ones that include objects or states of affairs that elicit specific kinds of responses. For instance, the category of the outrageous probably refers (roughly) to *that which elicits anger in normal observers*. Philosophers have long discussed the existence and metaphysics of response-dependent categories and debated about their role in moral evaluation (e.g. Gibbard 1992; J. Prinz 2007).

For my purposes, what is important about response-dependent categories like the outrageous is that one can judge that something falls under the category (or that something is outrageous), without experiencing or manifesting any anger or outrage. Not only is it likely that there is such a category, such a category would not develop were it not for experiences and observations of anger (in both oneself and in others). Thus, it makes sense that when considering a moral offense (as in the questionnaire studies above) one could judge that the offenses were more or less outrageous, without experiencing any actual outrage or anger. Moreover, insofar as anger directly influences punishment of wrongs to oneself (as in the UG), it may also influence how one thinks about outrageous actions directed at others. Specifically, it is easy to see how one could come to think of outrageous actions as *those toward which punishment and retaliation is appropriate*. If so, then we can imagine someone engaging in impersonal punishment because they judge an action to be morally outrageous, and because they judge that punishment is an appropriate response to morally outrageous action. If people make intuitive judgments in this way, then this sense of appropriateness could easily be rationalized as desert.

If this is correct, then we have a mechanism by which anger can influence cold or impersonal judgments and actions regarding punishment by influencing the development of the category of the outrageous. Moreover, this would explain why judgments of outrage and anger often accompany impersonal punishment even though we have reason to believe that such actions are not accompanied by affective arousal. Finally, it is a mechanism by which retributive motives (and the accompanying notions of appropriateness and desert) can be extended from the personal domain (as in the UG) to the impersonal domain (as in responses to questionnaire studies and punishment in other economic games).

Little evidence has been collected that would (dis)confirm this hypothesis. Moreover, it is unclear which competing hypotheses might also explain the data. There is some indirect evidence that childhood experiences with anger influence punitive judgments later in life. For instance, children whose parents practiced corporal punishment are more likely to affirm the death penalty as adults, and this effect is mediated by trait anger (M. Milburn et al. 1995; M. A. Milburn, Niwa, and Patterson 2014). This effect falls short of directly confirming my hypothesis about the development of the category of outrageousness, but it does suggest that a child's development influences their affective responses in a way that may also influence their punitive judgments. In any case, my hypothesis has a good deal of initial plausibility, and I will adopt it in what follows as a working hypothesis.

3. The phenomena of anger and the theory of basic emotions

The connection between anger and punishment raises the question of what anger is. Unfortunately, there is a surprising degree of divergence among researchers concerning anger's function and characteristics. It has been described as "a fairly specific syndrome (or network) of motoric, somatovisceral, and cognitive reactions…" that is caused by aversive stimuli and that includes an aggressive impulse (Berkowitz 2012a); as an appraisal of an action as "a demeaning offense against me and mine" (Lazarus 1991); as a type of emotional arousal that can lead to aggression or a number of other outcomes depending on social learning (Bandura 1973); as a product of social construction that bears a complex relationship with aggression (Averill 1983); as a capacity that functions "(a) to influence others to obtain some benefit, (b) to express grievances and establish justice, and (c) to assert or defend social identities" (Tedeschi 1994); as an evolved mechanism to regulate the dispositions of others toward oneself (A. N. Sell 2011); and as a mechanism for enforcing a specific kind of moral norm (Rozin, Lowery, and Haidt 1999).

While there are some important themes and commonalities shared between these theories and descriptions, debate continues concerning the nature of anger (see e.g. Berkowitz and Harmon-Jones 2004; Roseman 2004; A. Sell, Tooby, and Cosmides 2009). For the sake of clarity, some rely on stipulative definitions. For some theorists, something is properly called anger only if it is caused by a specific cognitive appraisal (Clore and Ortony 1993). On this view, an emotional state constitutes anger only if someone appraises an aversive event as intentionally caused by an agent. On the other hand, if an aversive event is appraised as unintentional or as lacking an agential cause, then it constitutes frustration. Others theorists defer to the prototypical structure of folk concepts of anger (Russell and Fehr 1994) in order to lump together the affective phenomena to which the word "anger" refers (Berkowitz 2012b). Evolutionary psychologists in the tradition of Cosmides and Tooby take a more theory-driven approach. They identify an adaptive problem: roughly, the need to get conspecifics to consider one's interests when acting, and they assume (without further argument) that anger is the solution to this problem (see e.g. A. Sell).

I cannot here offer a full criticism of these approaches. Suffice it to say that on each of these views, anger will end up looking like a conventional kind. Another example of a conventional kind is the culinary category of fruits (which is sometimes contrasted with the *biological* category of fruits). This is not a natural category because the difference between culinary fruits and vegetables is mind-dependent, reflecting a difference in the way that human palates respond to plant produce. Likewise, the characterizations of anger above reflect human stipulation, explanatory interests or linguistic practices rather than real divisions in nature. One problem with these approaches is that they tend to be arbitrary, sometimes ruling out possibilities by fiat. For example, the evolutionary psychologist's specification of anger rules out two possibilities. First, the adaptive function they specify could be implemented by *several*

underlying biological or psychological entities, only one of which might answer to the word "anger". Second, anger could implement this adaptive function along with several other distinct adaptive functions. It seems quite possible, given the tinkering nature of the evolutionary process, that what we call anger contributes to a number of different adaptive functions (or has contributed to a number of different adaptive functions during its evolutionary history) or that anger works synergistically with other psychological processes to produce adaptive behavior. If at the outset, one assumes or stipulates that anger is *the* faculty that solves a specific adaptive problem, then these possibilities are closed off without further justification.

My purposes demand further justification. For me, the claim defended in the previous section, that anger produces retributive intuitions, is an interesting empirical claim. But the more latitude one has to choose a definition of anger to suite her purposes, the more stipulative and the less empirical such a claim would become.

Fortunately, one need not be tempted in this manner. There are several clues that anger is not merely a conventional kind, the reference of which depends entirely on, say, human social practices, linguistic practices or explanatory interests (among other things).[14] As I describe in greater detail below, humans make involuntarily facial expressions, and even across cultures, there is general agreement about which facial expressions are expressions of anger and about what situations elicit those expressions. Even people who are born blind and deaf make involuntary facial expressions that the sighted can identify as expressions of anger. If anger were merely a conventional kind, then the development of these involuntary facial expression as well as their categorization would presumably depend on some kind of social transmission. In that case, we would not expect cross-cultural similarities in facial expressions and in the way people categorize them. Nor would we expect similarities in the situations that elicit a

[14] In chapter 3, I make the case that basic human anger is also a natural kind.

given expression (at least when there are boundaries across which cultural transmission does not occur). Moreover, we would expect the congenitally blind not to manifest facial expressions of anger. However, this is exactly what has been found concerning anger.

These phenomena are not restricted to anger, but there is a broader class of emotions for which they can be observed, the *basic emotions*. These phenomena provide several clues about the underlying systems that produce them and about the evolutionary forces that shaped those systems. In the paradigmatic cases of anger, fear, disgust, sadness, joy, and surprise, basic emotions have facial expressions that are recognized and produced across all cultures that have been observed (Ekman, Sorenson, and Friesen 1969; C E Izard 1994).[15] These expressions appear very early in development as does the capacity to recognize or respond to them (e.g. Carroll E. Izard, Hembree, and Huebner 1987). They appear spontaneously in those who could not have learned them from experience, as in those who are born both blind and deaf (Eibl-Eibesfeldt 1973; Eibl-Eibesfeldt 1979). They are automatic in the sense that they come unbidden and are difficult to fake, suppress or control.

Many of the facial expressions of basic emotions have similar expressions in chimpanzees and other non-human primates (e.g. Chevalier-Skolnikoff, 1973; Parr, Waller, Vick, & Bard, 2007). This means that the sets of muscle contractions involved in human facial expressions of these emotions naturally occur in the social interactions of these species. The facial expressions of basic emotions are also tied to distinctive changes in physiological arousal. For instance, distinctive changes in heart rate and blood pressure associated with anger can be activated in response to angering situations as well as by voluntarily contracting the facial muscles involved in anger expression (e.g. Levenson, Ekman, & Friesen, 1990). Recognizable changes in vocal characteristics have

[15] There may be other basic emotions that are not recognized via facial expressions but instead by their postural characteristics. Two prominent examples are pride and shame. See Clark (2009) for an argument that these emotions are basic.

also been found for each basic emotion, with anger being the most recognizable (Scherer, 2003).[16]

Less important for my purposes are the manner in which these emotional responses can be elicited. Basic emotion can be elicited quickly and automatically, below the threshold of consciousness and sometimes in a way that entirely bypasses higher perceptual processing (LeDoux 1998). The situations that elicit basic emotions often constitute universal themes (Ekman 1999). For instance, the elicitors of fear are more or less captured by the theme of danger (Öhman 1986) and the elicitors of disgust are more or less captured by the theme of offensiveness and contamination (Rozin and Fallon 1987). While these themes seem universal, cultural differences generate variations on these themes. There are also individual differences in elicitors within cultures, but some elicitors for a given emotion are easier to learn and more difficult to unlearn than others, with phobias (acrophobia, arachnaphobia, etc.) and taste aversions being prime examples. While none of these features are the *sine qua non* of basic emotions, all of the basic emotions that have been identified have most of these features.

Basic emotion theories are committed to explaining these phenomena in specific ways. Indeed, these features of basic emotions place important constraints on both their proximate and ultimate explanation. Different facts about basic emotions (or different phenomena) constrain their proximate explanation by implicating certain kinds of internal mechanisms. For instance, the *production facts* include the existence of universal signals, distinctive physiology, and coordination between the two. They also include other facts about facial expressions: that regular development of facial expressions can occur despite deprivation; that they are involuntary and difficulty to fake, suppress or control. These production facts about each basic emotion are thought to be explained by internal mechanisms called *affect programs*. Affect programs are behavior

[16] For an accessible overview of this body of research, see Ekman (2003).

programs which store "the patterns for these complex organized responses, and which when set off directs their occurrence." (Ekman, 1977)[17]

An underlying theme that unites these phenomena is that most of them (e.g. early development, universal signals) provide evidence that the idiosyncrasies of individual experience must play a limited role in shaping these emotional responses. For instance, early development of emotional expression and recognition suggests that these emotional capacities develop faster and with more regularity than they would if they relied upon general-purpose learning mechanisms in response to the information present in individual experiences. This (among other features of basic emotions) suggests that as the inheritors of basic emotions, human beings have more knowledge and know-how than our individual experience could afford us.[18]

When we ask what else, besides experience (and culture), might have shaped basic emotions, the best explanation is biological. Early on, basic emotion theorists suggested that basic emotions have genetic bases and dedicated neural substrates

[17] See Griffiths (1997, 88–91) for a nuanced discussion of the possible control mechanisms for affect programs. A more detailed sketch of the putative mechanisms that explain basic emotion phenomena would include the following. The *elicitation facts* (including fast, automatic, and unbidden occurrence) suggest that emotional responses are triggered by dedicated *automatic appraisal mechanisms*. The *limited variation facts* (including easy learning, difficult unlearning, and universal themes) suggest that appraisal mechanisms (automatic or otherwise) are calibrated by *prepared learning mechanisms*. The *recognition facts* such as early recognition and selective impairment of emotion recognition (e.g. Calder et al. 2004) may suggest domain-specific and perhaps emotion-specific *recognition mechanisms* that operate either empathically or on the basis of innate facial or postural templates. One set of phenomena, notable for its absence in the foregoing includes further *dissociation facts*, which suggest that for at least some basic emotions, some of the internal mechanisms mentioned (e.g. for elicitation, production and recognition) are *emotion-specific*. For example, Panksepp and Biven (2012) is a recent and accessible discussion of the dissociation facts and their implications.

[18] Culture is an important repository of knowledge that fills the gap between our sparse experience and our abundant knowledge (Richerson and Boyd 2004). Moreover, emotion is likely to be an important means by which cultural information is transmitted (e.g. Kelly 2011). But while cultural variation and transmission can explain many of the *differences* in emotion elicitors, cultural commonalities do not seem to explain *common themes* to which each basic emotion responds. It is more plausible to suppose that the explanatory arrow goes the other way, that shared emotional themes explain cultural commonalities. Emotions seem to develop in human infants before culture has fully taken root, thus emotions are also likely to explain the limits on the range of cultural variants observed in emotion elicitors. For instance, disgust seems to explain which etiquette norms persist over time (Nichols 2004, chap. 4), and it also seems to explain many of the cross-cultural features of purity norms (Kelly 2011, chap. 4). The general point is that basic emotion theory is consistent with social constructivist views of the emotions (see esp. R. Mallon and Stich 2000).

(Ekman, 1977). However, it is often difficult to characterize the exact sense in which traits are genetically determined (see e.g. Griffiths & Machery, 2008). This is because developmental processes that produce these traits often depend on environmental regularities (non-genetic determinants) for species-typical outcomes (cf. Griffiths, 2001; Ron Mallon & Weinberg, 2006). However these details are to be worked out, there is a clear sense in which basic emotions are *innate*. We have basic emotions because of our biological constitution and not because of environmental regularities.

The intellectual pedigree of basic emotion theory bears a close relationship with the ethological tradition of Konrad Lorenz. Not surprisingly, the notion of innateness employed within basic emotion theory closely resembles his concept of "phylogenetic information". Responding to the criticisms of developmental biologists in his later work, Lorenz allowed that there may not be any clear and useful distinctions between traits that are innate or acquired. Nonetheless, he believed that "there is a sound distinction to be drawn between…two sources of adaptive information." (Browne, 2005) One source of adaptive information comes from learning; this *ontogenetic* information is acquired, not given. Another source of adaptive information derives from evolutionary history; this *phylogenetic* information is given, not acquired. One important empirical method for demonstrating that a trait is shaped by phylogenetic information is through deprivation experiments. By depriving an organism of relevant experiences (e.g. depriving a bird of exposure to its species-typical bird song), one can sometimes show that the organism nonetheless retains a specific form of adaptedness to its environment (e.g. if after deprivation, the bird is nonetheless capable of singing its species-typical bird song). For example, when children are born blind, they are deprived of visual experiences that would be relevant to the acquisition of normal facial expressions of anger through learning. Nonetheless, these children display species-typical anger expressions in the appropriate contexts. In Lorenz's terms, facial expressions of human anger are best

explained by the possession of phylogenetic information and are thus innate adaptations.[19]

Of course, the idea of phylogenetic information cannot be understood without the concept of adaptedness. Accordingly, there is broad agreement among basic emotion theorists concerning the adaptedness of basic emotions. In harmony with Lorenz, the adaptive fit between emotions and the relevant environments is supposed to be best explained by our ancestry and the selective pressures that operated on it: "…emotions evolved for their adaptive value in dealing with *fundamental life tasks*." (Ekman, 1999) These fundamental life tasks include avoiding poisons, parasites, and predators as well as dealing with gains, losses, and resource competition.[20] These situations are thought by many to constitute "…commonly recurrent (across generations) adaptive situations." (Tooby & Cosmides, 1990, pp. 407–408) Basic emotions were shaped by these recurrent situations, endowing us with knowledge or know-how that is given (by our ancestry) and not acquired (by individual experience).

The linkage of basic emotions to facial expressions suggests that their fundamental life tasks all have important *interpersonal* dimensions. Ekman, for instance, thinks that "…the primary function of emotion is to mobilize the organism to deal quickly with important *interpersonal* encounters, prepared to do so by what types of activity

[19] While I mentioned above some of the controversies surrounding the concept of innateness, an interesting proposal has been put forward recently (O'Neill) that I think nicely captures the sense of innateness employed by Lorenz and others. According to this account, innateness is the insensitivity of the development of a trait to specified environmental variation. This view proposes a terminological corrective to claims of innateness. These claims should never be made in absolute terms, rather claims of innateness are relative to specified environmental variation. In the present case, facial expressions of human anger are innate *with respect to* deprivation of visual and auditory information. One might also interpret the early cross-cultural experiments evaluating the recognition of basic emotions (Ekman et al., 1969; C. E. Izard, 1971) as substantiating a claim of innateness *with respect to* cultural variation.

[20] As LeDoux (1998) points out, solving fundamental life tasks requires very different kinds of functions. Thus, he suspects that "…there must be different brain systems to take care of these different kinds of functions." (p. 126) So in what follows, I will not assume that there is a single automatic appraisal mechanism or affect program that functions to produce the variety of basic emotions. Instead, I will call the affect programs for different emotions by different names (the *anger* affect program as opposed to the *fear* affect program), but acknowledge the possibility that we may eventually discover some of them to refer to the same brain systems. For instance, Gray (2003) might be taken to suggest that anger and fear are the product of a single system.

have been adaptive in the past." (Ekman, 1999) Insofar as basic emotions aid in the avoidance of poisons, parasites and predators and help to deal with gains, losses and resource competitions there is likely to be an interpersonal dimension to all these tasks. We avoid poisons and parasites together, and we negotiate resource competition primarily with other members of our species. If this is right, it makes sense that basic emotions are tied to our biological constitution. As Ernst Mayr notes, "Since much of the behavior directed toward other conspecific individuals consists of formal signals and of appropriate responses to signals, and since there is a high selective premium for these signals to be unmistakable, the essential components of the phenotype of such signals must show low variability and must be largely controlled genetically." (Mayr, 1974, p. 657)

In sum, we can condense basic emotion theory into the following claims about basic emotions: basic emotional responses are *innate adaptations* that contain *phylogenetic information* for dealing with *recurrent, interpersonal situations* in our selection history. Moreover, these adaptations include affect programs that coordinate various response components of a given emotion syndrome. At least some of the psychological states answering to the term "anger" appear to be instances of a basic emotion.

As well established as this theoretical framework is in some circles, there have been influential criticisms of it (Barrett 2006). So it is worth pointing out that my view is consistent with some data that conflict with basic emotion theory (according to some). First, while I *am* committed to there being an evolved anger syndrome in some humans that coordinates automatic facial expressions and physiological changes, (perhaps also postures and motivations), I am not necessarily committed to the claim that each symptom of the syndrome (physiological changes, production of facial expressions,

recognition or discrimination of facial expressions etc.) will be regularly manifest within individuals, across individuals or across cultures.

By contrast, some basic emotion theorists are committed to defending the universal recognition of certain facial expressions. Nevertheless, the widespread recognition of certain facial expressions is only a symptom of something else, namely that *in some environments*, an emotion syndrome can produce visible changes in facial expression, and that in those environments, people can become consciously aware of the similarity of those expressions (or their difference from other expressions). For instance the Dani people of West Papua cannot reliably distinguish between facial expressions of disgust and anger. Nevertheless, the ability to discriminate the two expressions would just be a symptom of the existence of two distinct underlying syndromes. This symptom could easily be eliminated if a culture had strong norms against overt confrontation or against the expression of anger. This would prevent people from becoming consciously aware of the difference between the two expressions even if there were real differences between them at the level of production. So the fact that some cultures cannot distinguish disgust and anger does not provide direct evidence against the claim that these two distinct evolved emotion syndromes exist. For similar reasons, it is not clear that we should expect to find recognition, discrimination or even distinct physiological changes for each basic emotion across all cultures or individuals.

In many cases, universality itself is only a symptom that some phenomenon is shaped or constrained by inheritance (though it could be a symptom of other things as well). Even if it lacked true universality, anger might still exist in many cultures as an evolved emotion syndrome shaped by inheritance. This is because there could be populations in which the syndrome has been weeded out or lost through genetic drift. Universality is a symptom of shared inheritance, but so is near-universality.

Variation is possible in many dimensions of an emotion syndrome across cultures, across persons and even within persons. This is partly due to distinctively human capacities for emotion regulation – which are highly flexible and distinct from basic emotion systems. As a result of these capacities, different individuals can regulate their emotions differently, and the same individual can regulate emotions differently at different times or under different conditions. Similar things might be said for capacities to regulate physiological arousal and to habitually inhibit unbidden facial expressions.

The existence of an evolved syndrome is also consistent with a plurality of psychological states, all of which answer to the term "anger". This could happen if, for instance, states of anger constitute serial homologies (e.g. Clark 2009). Serial homologies are repeated structures within an individual organism the classic example of which are the vertebrae. Each vertebra is an instance of the same original structure, repeated within the same organism to comprise the organism's backbone. There has been some suggestion that psychological structures, like emotions, could also be repeated within an organism. As a result, we could imagine a person with several different psychological states that all have a similar structure. For instance, we could imagine how a state of righteous indignation and a state of childish frustration could be distinct, but could also motivate similar behaviors, for example retribution and reactive aggression, respectively. These might be different psychological states with different underlying mechanisms and different elicitors (e.g. moral violation as opposed to removal of a reward contingency), and yet have a highly similar motivational structure (e.g. cost imposition in reaction to provocation) due to their derivation from a common preexisting structure. The concept of serial homology shows how it is possible to have a number of distinct psychological states that all answer to the term "anger". In conclusion, an evolved anger syndrome is consistent with variation in manifestation (within

individuals and across cultures and individuals) and is consistent with multiple forms of anger.

4. Chapter summaries

In the course of this dissertation, I bring together numerous strands of research in order to begin a natural history of anger and punishment. The overall argument is that anger is not unique to humans and was shaped largely for its role in negotiating resource competition (in an era of adaptedness shared with non-human animals), as were the retributive motives produced by anger. Moreover, this natural history has clear normative implications for the justification of punishment.

In chapter 1, I begin by making the case that, at least in principle, the natural history of anger can undermine the role of retributive considerations in justifying normative theories of punishment. At the heart of this argument is a simplified "just so" story about the evolution of retributive motives. Nevertheless, this story is sufficient to animate the in-principal argument against retributive considerations. Moreover, this in-principle argument may generalize to other non-consequentialist considerations, ones that favor actions like keeping promises, cooperating with others, and avoiding intentional harm to others.

In chapter 2, I begin moving from a "just so" story toward a more methodologically rigorous explanation of retributive motives. I appeal to a selection model for resource competition to explain how retributive motives could have arisen in organisms that lack the capacity for complex, strategic social interaction. As such, the selection model applies (if at all) to events very far back in our evolutionary history, prior to the evolution of complex social organization. Nevertheless, instead of tying this explanation directly to human evolution, I show that this selection model explains the structure of one pattern of aggressive behavior in rats and rodents more broadly, which

has been carefully characterized through observation and experiment over the past several decades.

In chapter 3, I further flesh out this explanation of retributive motives by arguing for a homology between human anger and the system responsible for this pattern of rodent aggression. This means that the rodent aggression system and human anger derive from a trait of the last common ancestor of humans and rodents. Along the way, I develop evidential criteria and constraints for adjudicating competing homology claims.

Finally, in chapter 4, I forestall a set of serious misgivings about this approach and pave the way for further developing the natural history of anger and punishment. These misgivings arise out of the fact that there are vast differences between rats and humans, and that some of these differences reside in the patterns of behavior that anger can give rise to. Whereas anger causes highly stereotyped aggression in rodents, human anger leads to highly flexible behaviors, only some of which include physical aggression. I argue that despite appearances, rat aggression is actually highly flexible and cannot be explained without appealing to an angry motivational state that is integrated with internal representational states of individual rats. I conclude by conjecturing at how the relatively primitive retributive motives of our common ancestor with rats could have been shaped by the process of evolution to be sensitive to more complex social aims and to produce highly variable behaviors manifested by angry people today.

Chapter 1
The Evolution of Retribution: Intuitions Undermined

Work in the last decade of empirical moral psychology suggests that emotions are responsible for at least some *deontological moral intuitions*. These intuitions are revealed by widespread tendencies to judge or act[1] contrary to a consequentialist evaluation of actions.[2] Deontological intuitions are thus a primary source of evidence against some consequentialist theories. Nevertheless, if emotions are responsible for these intuitions, as empirical studies suggest, then there is reason to reconsider their role in philosophical theorizing.[3] A prominent criticism of deontological intuitions, offered by Joshua Greene and Peter Singer, is that empirical moral psychology reveals epistemic defects in the emotions responsible for deontological intuitions.[4] On this view, consequentialism is vindicated because one important source of evidence against it is not good evidence after all.

These criticisms of deontological intuitions include evolutionary debunking arguments. For instance, one of Joshua Greene's criticisms is this: emotions were

---

[1] I intend intuitions to include non-inferential inclinations to judge a proposition only by considering its content and non-inferential inclinations (not) to perform an action only by considering some representation of the action or situation. Cf. (Sosa 2007, 233; Sinnott-Armstrong 2008, 209) The non-inferential nature of intuitions refers to the fact that one can have a feeling that something is right or wrong without any accompanying justificatory explanation for the feeling. On my view, "intuition" is a theoretical term capturing a set of phenomena with common explanatory elements (perhaps they are all caused by some psychological process or another) that are explananda or objects of study in scientific fields like philosophy, moral psychology, behavioral economics and social psychology. *Deontological intuitions* refer to a subset of these phenomena that has a common underlying psychological cause. Accordingly, one can have a deontological intuition without judging its content true or its practical conclusion prudent or morally right, but I do not think anything hinges on this terminological decision. Some philosophers argue that intuitions are properly understood as judgments. Whether or not there exists phenomena involving *inclinations* to judge or act (regardless of whether they in fact lead to judgment or action) does not depend on how philosophical debate proceeds. If the phenomena exist, then these phenomena are a proper object of study regardless of whether they consistently give rise to actual judgments.

[2] More specifically, they are tendencies to judge or act contrary to an act-consequentialist decision procedure: "On each occasion, an agent should decide what to do by calculating which act would produce the most good." (Hooker RRR). Since this definition of deontological intuitions pits them directly against consequentialism, I use "deontological intuition" and "anti-consequentialist intuition" interchangeably. Thanks to [name revoked] for pointing out tendency of empirical moral psychologists to conflate consequentialist decision procedures and consequentialist standards of rightness.

[3] See e.g. the introduction of Kamm, (1993).

[4] (Singer 2005; J. D. Greene 2008)

selected for their role in increasing fitness and deontological intuitions are a byproduct of this evolutionary function, thus emotions would have produced deontological intuitions whether or not these intuitions were true. The most common objection to this argument is that it proves too much. It threatens to undercut a wider set of evaluative intuitions, ones that share the same kind of evolutionary explanation and some of which support consequentialism and evaluative realism.[5] This is a problematic feature of Greene's and Singer's evolutionary debunking argument, because it vitiates their aim of defending consequentialism. Moreover, there is a suspicion that similar problems will plague any evolutionary debunking argument pitched at the level of first-order moral discourse.[6]

I believe that this is not an essential feature of evolutionary debunking arguments. Thus, my purpose here is to give an evolutionary debunking argument against deontological intuitions (or at least a subset of them) that avoids this charge and therefore applies directly to normative ethics, undermining only deontological intuitions. To make this argument, I propose a friendly amendment to Greene's dual process account of moral intuition. Rather than capturing the difference between deontological and consequentialist intuitions in terms of the difference between emotion and cognition (as do Greene and Singer), I capture the difference with a distinction between *prospective* and *non-prospective* processes. Non-prospective processes place *non-derivative* value on actions (or action types), value that does not derive from the action's consequences. Anger is an example of a non-prospective process, because it places non-derivative value on actions of revenge and retribution.

Once this distinction is in place, I present a novel debunking argument. If anger produces retributive intuitions (which are one species of deontological intuition) because

[5] (See e.g. Berker 2009; Kahane 2011; Mason 2011)
[6] (E.g. Kahane 2011; Mason 2011; Vavova 2014)

of the biological consequences (e.g. increased fitness) of those intuitions, then the intuitions are not good indicators of non-derivative value. This severs the putative evidential connection between retributive intuitions and the non-derivative, or retributive, value of punishment. Thus, retributive intuitions are not good evidence for retributive theories of punishment (according to which punishment has non-derivative value).

1. Greene's Evolutionary Debunking Argument

1.1 Debunking arguments

I begin by briefly characterizing debunking arguments. Debunking arguments attempt to undermine beliefs, intuitions, values or judgments by impugning their causal source. In their debunking arguments, Greene and Singer evaluate the putative evolutionary causes of deontological intuitions in addition to their immediate psychological causes. For an example of the latter kind of debunking argument, consider the moral intuitions involved in two classic moral dilemmas. In the *trolley* scenario, a trolley car is hurtling toward five unsuspecting workers and one faces the option of flipping a switch to divert the trolley to a track with a single worker. In the *footbridge* scenario, the trolley poses the same threat to five workers and one faces the option of pushing a large man off a footbridge and in front of the trolley to stop it in its tracks. There is an intuitive difference between hitting a switch to save the five and pushing a man off a bridge to save the five. Moreover, this difference has been taken to support moral principles like the doctrine of double effect.[7] However, one might think that if these intuitions "reflect the influence of morally irrelevant factors…[then they are] unlikely to track the moral truth…".[8] For instance, Greene and colleagues have argued that the intuitive difference between the cases is best accounted for by the exertion of

[7] Foot makes this argument in (1978)
[8] (J. D. Greene 2008, 70)

personal force with intention to harm in *footbridge* and by its notable absence in *trolley*.[9] Greene hypothesizes that the response to *footbridge* is thus driven by an aversive emotional response that is triggered by these factors. To Greene and others, whether cases differ along some of these dimensions (e.g. personal/impersonal) seems morally irrelevant, so he claims that these emotion-driven intuitions do not track the truth and thus do not count as evidence against consequentialism.[10]

While Greene talks about "truth tracking" and "morally irrelevant factors" to describe the epistemic defect in deontological intuitions, I will instead cast debunking in terms of evidential defeat. A piece of evidence rarely provides conclusive support for a conclusion because evidence can sometimes be defeated or overturned.[11] For example, one might have evidence that a painting is a Monet because of the report of an art dealer. This evidence could be defeated in two ways. If I received a conflicting report from a museum curator, whose authority is unsurpassed, then this evidence would outweigh my other evidence. It would thus constitute a *rebutting defeater*, because it outweighs my other evidence by giving me stronger reason to deny that the painting is a Monet. Alternatively, if I received word that the art dealer had lied about the provenance of several paintings in the past (not including the putative Monet), this could undercut my other evidence entirely, constituting an *undercutting defeater*.[12] Rather than providing evidence against the art dealer's claim, it severs the evidential connection between the art dealer's report and the state of affairs that it reports. In other words, it gives me reason to think that the dealer's report is not a good indicator of the state of affairs that it reports. While the undercutting defeater does not give me reason to believe that the art piece is not a Monet, it does eliminate my primary reason for thinking that it is. I find it

[9] (e.g. J. D. Greene et al. 2009)
[10] Though Greene criticizes these intuitions in a number of other ways (see e.g. J. Greene 2003; J. D. Greene 2008).
[11] (see e.g. J. L. Pollock 1987)
[12] Pollock (e.g. J. Pollock 1986) was the first to recognize the distinction between undercutting and rebutting defeaters.

natural to say of this case that the evidence of the art dealer's dishonesty debunks my evidence that the painting is a Monet. So I propose to understand debunking arguments in terms of undercutting defeat.

While this example is framed in terms of a person's testimony and the states of affairs it reports, we might just as well apply these evidential considerations to psychological processes and the states of affairs they represent. For instance, we can imagine that the psychological processes that produce deontological intuitions are in some way disconnected from the states of affairs that they "report" (via the intuitions they produce). If one came to know about this disconnect, then one should not rely on deontological intuitions and should not include them in one's evidence base. In that case, one would have an undercutting defeater for deontological intuitions or equivalently, one's deontological intuitions would be debunked. While I will present Greene's argument in his own terms (e.g. truth-tracking), later I will cast my own argument in terms of severed evidential connections and undercutting defeat.

1.2 An evolutionary debunking explanation for emotional processes

Greene proposes an evolutionary explanation for the differences in moral judgment between cases like *trolley* and *footbridge*:

> The emotions most relevant to morality exist because they motivate behaviors that help individuals spread copies of the genes they possess within a social context… [Evolutionary] theories explain the widespread human tendency to engage in cooperative behaviors (e.g., helping others and speaking honestly) and to avoid uncooperative behaviors (e.g., hurting others and lying), even when relatives and close associates are not involved…
>
> I will simply assume that the general thrust of these theories is correct: that our basic moral dispositions [which are motivated by moral emotions] are evolutionary

adaptations that arose in response to the demands and opportunities created by social life.[13]

Here, Greene is borrowing from evolutionary theories that attempt to explain human altruism. These explanations highlight situations in which altruism contributes to an organism's ability to spread copies of its genes or to a group's ability to survive or multiply.[14] These theories go some distance in explaining why we have evolved dispositions to avoid hurting others, *inter alia* (e.g. dispositions for punishing, keeping promises or for helping others).

Moreover, there is a reason why these dispositions would only manifest in cases like *footbridge*. Opportunities to harm others in distant and detached ways, as in *trolley,* were not available to our ancestors, and Greene thinks that this is a "…contingent, nonmoral feature of our evolutionary history."[15] Finally, there is a reason why emotions rather than slow, effortful, deliberative processes would implement these dispositions. Emotions are "fast and frugal". They are a simple and efficient way of responding to the types of recurrent situations that they were selected to deal with.[16] They are elicited by a small range of factors (e.g. that an action involves personal force) that are nonetheless reliable indicators of the relevant situation, and they result in adaptive inclinations to behave or judge (e.g. against hurting and toward helping). Greene concludes from this evolutionary genealogy that deontological intuitions do not track the truth.

1.3 The evolutionary debunking argument proves too much (or nothing at all)

[13] (J. D. Greene 2008, 59–60)
[14] For instance, a kin selection explanation of altruism appeals to the benefits that accrue to helping one's kin, even at a cost to oneself. Since kin are more likely to share copies of the same genes and since kin-helping can facilitate the spread of those copies, genes that induce altruism toward kin can spread in a population under certain conditions. See (Hamilton 1964) Other evolutionary explanations attempt to identify selective benefits of altruistic behavior in a much larger range of cases, for instance some appeal to the selective benefits of living in larger groups and transmitting knowledge through culture. See (Richerson and Boyd 2004; S. Bowles and Gintis 2004).
[15] (J. D. Greene 2008, 70)
[16] (Tooby and Cosmides 1990; Ekman 1999)

However, one of the premises in Greene's argument is ambiguous. What exactly is the morally irrelevant factor introduced by the evolution of alarm-like emotions? It could be that emotions evolved *to respond to* a morally irrelevant factor, namely personal force. Alternatively, it could be that emotions evolved *in response to* a morally irrelevant factor, namely that deontological moral inclinations helped our ancestors to spread their genes. If we opt for the former, then it is unclear what role evolution plays in this argument. If one already knows that it is morally irrelevant whether an action requires personal force and if one already knows that this is what makes *footbridge* and *trolley* seem different, then we already have a debunking explanation of the intuitive difference and evolution adds nothing. On the other hand, if one has good reason to believe that personal force is relevant to moral evaluation, then the fact that emotions evolved to respond to them should only vindicate the emotional response.

If evolution is doing any additional work, it must be showing that the processes responsible for these intuitions evolved *in response to* morally irrelevant factors: that deontological inclinations helped our ancestors spread their genes and not that deontological intuitions are true. In other words, the function of evolution must be to convince us of something like the following: for all we know, if deontology were false, evolution would shape our emotions to give us the same stock of deontological intuitions.

Nevertheless, this kind of evolutionary consideration could easily undermine intuitions that support consequentialist theories. Consider the fact that consequentialism is vacuous without a theory of value. Without some idea of what outcomes should be valued or disvalued, there would be no way to determine which action would be right by the lights of a consequentialist moral theory. One would be unable to evaluate the consequences of actions. Moreover, intuitions about what things are valuable will inevitably influence any theory of value. This makes consequentialism vulnerable to a well known evolutionary critique. For instance, if the arguments of Sharon Street are

correct, then we hold many if not all of our evaluative intuitions not because they are true (in the sense required by realist theories of value) but because they allowed our ancestors to more effectively spread their genes.[17] This applies to even the most uncontroversial of evaluative intuitions: that pain is bad, that it is bad to hurt others, and that it is good to help them. These arguments are supposed to result in a more global form of evaluative skepticism than Greene wants; one that vitiates any value theory on which his consequentialism might draw.[18]

If we can undermine deontological intuitions just by pointing out this kind of evolutionary influence, then a similar conclusion should follow in the case of evaluative intuitions more broadly. Therefore, Greene's evolutionary considerations threaten to debunk not only deontological intuitions but also the intuitions that ground any consequentialist theory of value.[19] The challenge for an argument like Greene's is to point out evolutionary considerations that undercut deontological intuitions without undercutting a much broader range of evaluative intuitions. This is what I aim to do in the following section.

2. Debunking Vindicated

2.1 How best to understand the dual process framework: reaction versus prospection

[17] (Street 2006)

[18] Things get slightly murky at this point because Greene is actually a moral antirealist of some kind. While his arguments seem targeted directly at moral realists, perhaps Greene really only wants to defend an antirealist version of consequentialism against deontological intuitions. In fact, he does briefly criticize what he calls "anthropocentric morality" a moniker meant to encompass antirealist and constructivist metaethical theories that might maintain deontological intuitions despite their genealogy (E.g. Korsgaard 1996). He argues that even from these metaethical standpoints, one should not count these intuitions as evidence if they are influenced by morally irrelevant factors. This argument turns on the same ambiguity as above. Either the morally irrelevant factors concern what we evolved to respond to or they concern what we evolved in response to. If it is the former, then evolution adds nothing to the argument. If it is the latter, then it is hard to see how evolutionary considerations do not vitiate the evaluative intuitions that are the basis for an anti-realist consequentialist's theory of value if they really do vitiate the deontological intuitions that are the basis for an anti-realist or constructivist deontology.

[19] Similar points have been made elsewhere. See Vavova (2014), Kahane (2011), Mason (2011) and Berker (2009).

A central part of Greene and Singer's debunking arguments is Greene's dual process theory of moral intuition according to which consequentialist and deontological intuitions have distinct psychological underpinnings.[20] The distinct evolutionary etiologies of these psychological processes are supposed to distinguish the epistemic value of the intuitions they produce. Specifically, the evolutionary history of the emotional processes responsible for deontological intuitions, is supposed to undercut these intuitions in a way that the evolutionary history of consequentialist processes does not.[21] Nevertheless, the preceding arguments suggest that a successful evolutionary debunking of deontological intuitions also requires distinguishing deontological processes from the broader class of evaluative processes. Below, I make a friendly amendment to Greene's dual process theory, one that identifies a distinguishing feature of processes responsible for deontological intuitions. However, only later (in section 3.2) will I show that it adequately distinguishes these processes from other evaluative processes in terms of their epistemic value.

The main contrast Greene employs is between cognitive and emotional processes. Emotional processes have what Greene calls "behavioral valence", meaning that these "alarm like" emotions have the following properties. First, they include inclinations to behave in specific ways or to judge those behaviors as appropriate. Second, they are elicited in response to a limited range of factors (such as the presence/absence of personal force), and finally, once triggered they can override cognitive processes. By contrast, cognition involves slow, flexible, and controlled processes.[22] Cognition aligns with consequentialism because both are "systematic and aggregative": "...the advantage

[20] The dual process view has actually been the most successful aspect of Greene's research program. There is a wide range of evidence that there is indeed more than one (though perhaps also more than two) processes responsible for moral intuitions. However, there remains some debate about what distinguishes the two processes and whether a division of the processes aligns with the distinction between consequentialist and deontological intuitions. See (Cushman, Young, and Greene 2007; Cushman 2013; Kahane et al. 2012; Paxton, Bruni, and Greene 2013; Kahane 2012).

[21] (Cf. Singer 2005, 350)

[22] (See e.g. Evans 2003; Stanovich 2004)

of having such neutral representations is that they can be mixed and matched…without pulling the agent in multiple behavioral directions at once…"[23] In other words, these representations are supposed to make possible the kind of systematicity characteristic of consequentialism, because they can take all of the consequences of an action into account. This is the most important contrast for Greene: cognitive processes can consider an indefinite number of different factors when deciding how to act, whereas emotional responses are triggered by only a few kinds of factors.

While Greene believes that there is a natural mapping between the content of consequentialism and the properties of cognitive processes and "between the content of deontological philosophy and the functional properties of [emotional responses]…",[24] these distinctions do not align. First, it is possible to imagine processes that consider multiple factors, but that are not fully consequentialist in their deliberations. For instance, retributivism about punishment is a deontological theory since a retributive justification of punishment refers to what a transgressor deserves based on what she did in the past rather than on the good outcomes that would attend punishment.[25] Nevertheless, moral agents can weigh consideration of desert against other considerations to yield an all-things-considered judgment. For instance, I might be motivated to punish a person because of what she deserves, but I might nonetheless adjust the severity of the punishment in relation to the consequences of punishing. In that case, I have a backward-looking, non-consequentialist motive for punishment that aggregates with other factors by a process that can consider whatever factors seem

[23] (J. D. Greene 2008, 64)
[24] (J. D. Greene 2008, 63)
[25] I suspect that this remains true even if there is a valued outcome, perhaps justice, that attends deserved punishment. To me, it seems unlikely that proponents of such a value would promote it rather than respecting it. That is, if one believes that it is intrinsically good to punish the deserving for past transgressions, then it seems inconsistent to then say that one should sometimes withhold punishment for the sake bringing about a greater quantity of deserved punishment. If my primary aim in an act of punishment is giving Jones her just deserts for a specific transgression, then the fact that this will cause Jones' brother to withhold punishment from three other deserving transgressors does not plausibly give me a reason not to punish Jones.

relevant to the question at hand.[26] Thus, what is distinctive about the psychological processes that consequentialism maps onto is not that they consider or weigh multiple factors (since they could easily consider non-consequentialist factors as well), but rather that they are *prospective*, or outcome based. That is, they place value on actions according to their anticipated, internally represented, outcomes (relative to the agent) and decide what to do only based on (positively or negatively) valued outcomes.

Notably, work in animal behavior and computational neuroscience has revealed neural systems in human and non-human animals that place value on actions with reference to a *causal model* relating the action to its outcome.[27] In contingency learning experiments, a rat learns not (only) that pressing a bar is a good *kind of action* to perform, but that pressing the bar produces a specific *outcome*, namely the delivery of a certain kind of food.[28] If the rat is satiated with (or conditioned to aver) that kind of food, or if pressing the bar ceases to deliver it, the rat will diminish its bar pressing behavior. One of the best explanation of these patterns is that the rat develops a causal model of the outcome of pressing the bar and changes its actions when the hedonic outcome is less favorable or when the causal model updates to include a relevant change in contingency.

A second problem with Greene's alignment of distinctions is that we can imagine there being emotion-like responses that are only sensitive to one or two factors, but that nonetheless line up with consequentialist rationales. For instance, one could design a robot with an alarm-like response to the detection of a doomsday device. When triggered this response would do whatever is required to destroy the doomsday device. Moreover, the emotion-like response need not take anything else into account (besides the

[26] See Kahane ((2012, 531–533)) for a similar argument that deontological reasoning often requires weighing different duties against one another. This too seems like a process that can take many factors into account. Though he does not draw the same conclusions that I do concerning prospective processes.
[27] Though it is not clear to me that these models are essentially causal. Perhaps in humans they can represent other assymetrical dependency relations such as the *in virtue of* relation or other non-causal explanatory relations such as those involved in mathematical explanations. Relations of this sort seem crucial for comparing the value of different possible worlds.
[28] (E.g. Balleine and Dickinson 1998)

existence of an armed doomsday device) to accord with a purely consequentialist judgment about what to do. Thus, being triggered in response to a few kinds of factors is not a distinctive feature of processes that produce deontological intuitions. Rather, I think the kinds of processes that map onto deontological intuitions, if any, are *non-prospective* processes.

Non-prospective processes form a disjunctive class because there are many ways to motivate action and moral judgment that are not prospective. For instance, some processes place value on actions according to past experiences involving actions *of that kind*. In a recent experiment, participants were asked to physically enact simulations of harmful actions (e.g. holding a fake gun to another person's head and pulling the trigger).[29] Physiological indicators of aversion prior to performing these actions were greater than when participants performed nearly identical actions (e.g. holding a spray bottle in the air and pulling its lever) and greater than when they observed other people physically simulating the same harmful actions. The aversive reaction to these actions is not readily understood as an aversion to their consequences (e.g. the pain or discomfort of the "victim"). Rather the better explanation is that participants had a stronger aversion to performing actions of a certain *type*, namely ones that resemble taking someone's life by putting a gun to their head and pulling the trigger.

Other non-prospective processes are *reactive*, in the sense that the aim of action (or the reason for which it is selected) is represented in relation to past or present occurrences (as opposed to internally represented future outcomes). While such a process may be directed at a future outcome in some external sense (e.g. directed at the outcome by design), it is not guided by an internal representation of a future outcome. For such a process, "…the orientation towards a future state…can merely involve change from the present, – change from now: disappearance of pain, disappearance of the

[29] (Cushman et al. 2012)

desired object being out of reach."[30] For instance, a heat seeking missile need not be guided by an internal representation of its target or of the "desired" outcome of hitting its target. Of the many ways one could design such a missile, one of the simplest would be for it to receive feedback signals that indicate whether a heat source is reducing or increasing its distance in a given direction. When appropriately connected to its controls, this feedback can guide the missile to its target. Again, one need not include in the missile's program any internal representation of its aim (hitting a moving object) or its physical target (such as the geometrical structure or distinctive color patterns of the target or its spatial location relative to other objects). It only requires feedback signals that adjust its path in reaction to the path of its physical target and that direct it toward the achievement of its aim.

There is evidence from animal behavior, neuroscience and psychology that prospective and non-prospective systems operate independently of each other (and sometimes antagonistically) in a range of animal species including humans.[31] Moreover, it is likely that in many of Greene's examples reactive processes are producing the empirical results that Greene attributes to alarm-like emotions.[32]

2.2 An evolutionary function of one reactive process

With the distinction between prospective and non-prospective processes in hand, I can now articulate a new debunking argument. If some non-prospective processes were selected for their fitness enhancing consequences but also cause deontological intuition because of these consequences, then the function of non-prospective processes

[30] (Frijda 2010, 572)

[31] I cannot review this evidence here, and in any case, much of it has been thoroughly reviewed elsewhere, see Cushman (2013) and Crockett (2013). While Cushman focuses on the distinction between two learning systems, he does not mean to exclude other kinds of non-prospective action selection mechanisms (personal communication). The psychologist Nico Frijda has long emphasized the importance of impulsive motivation, which has precisely the characteristics of the non-prospective processes that I discuss. See (Frijda 1986; Frijda 2010)

[32] While space does not permit a thorough demonstration of this claim, I will consider Greene's example concerning retributive punishment in detail below.

(producing good outcomes) disconnects them from the states of affairs that they report (that actions have value aside from their outcomes). Thus, the evolution of non-prospective processes gives us an undercutting defeater for deontological intuitions. In this section, I sketch out this argument with reference to a specific reactive process, namely anger.

Consider a suggestion about anger similar to those made by Green, Singer and even Parfit[33]: anger is responsible for the intuitions that support an anti-consequentialist, *retributive* principle.

> R – The value (or justification) of an act of punishment is not (or not only) derived from the consequences of punishment.[34]

Greene appeals to evolutionary accounts of altruism to explain why an emotional response like anger would lead to the intuitions that support R: "…we have a taste for retribution, not because wrongdoers truly deserve to be punished regardless of the costs and benefits, but because retributive dispositions are an efficient way of inducing behavior that allows individuals living in social groups to more effectively spread their genes."[35] Here as elsewhere, Greene is appealing to evolutionary considerations of the sort that would also undermine the theories of value that consequentialism requires. For the argument to work in favor of consequentialism, we need a slightly different story about why the evolutionary function of anger makes it untrustworthy with respect to R.

To understand this function, it helps to get oneself in the grip of a puzzle. It is obvious that humans cooperate in a wide range of circumstances. There is cooperation not only among people who are genetically similar (a phenomenon that the theory of kin

[33] (Parfit 2011, chap. 429)
[34] Thanks to an anonymous referee for urging me to clarify a previous formulation. There is some psychological evidence for intuitions that support this principle (e.g. K. M. Carlsmith, Darley, and Robinson 2002b; K. Carlsmith and Darley 2008). If this and related research is not entirely convincing on this count, the reader can take this claim as a hypothetical assumption to demonstrate the validity (if not soundness) of the debunking argument.
[35] (J. D. Greene 2008, 71)

selection explains) and not only among people that frequently interact (a phenomenon that theories of direct reciprocity explain) and not only among people who signal an intention to reciprocate (a phenomenon that indirect reciprocity and costly signaling explain) but also, puzzlingly, "...among genetically unrelated people, in non-repeated interactions, when gains from reputation are small or absent."[36] This last kind of cooperation is beneficial because it may have allowed our ancestors to live in larger groups and receive the benefits of doing so.[37] Nevertheless, it remains controversial how we evolved the tendency to do so. Punishment could help explain this phenomenon, because it can help to secure cooperation. Nevertheless, punishment is costly. Even granting that the group level benefits of punishment (the ones that accrue to individuals in large cooperatives) outweigh the immediate costs,[38] what would motivate relatively short-sighted individuals to consistently punish in the conditions characteristic of life in large groups (non-iterative interactions amongst genetically heterogeneous individuals where reputation is difficult to track)? What would make someone willing to commit to punishment in the face of its momentary costs?

Fehr and Gächter set out to answer these questions (among others) with an economic game.[39] They set up the purest instance of the kind of interactions that we are concerned with (ones that are non-repeating and anonymous) by having groups of four people (anonymous to each other) play a "public goods" game. They gave participants an endowment of 20 monetary units (MUs) and then gave them the opportunity to invest it in a group project. The group project would return .4 MUs to each group member for every one MU invested, and at the end of the round, each player received information about how much other group members donated. To see how participants changed their strategy over time and with changing conditions, Fehr and Gächter had participants play

[36] (Fehr and Gächter 2002, 137)
[37] (E.g. Richerson and Boyd 2004)
[38] (E.g. Robert Boyd, Gintis, and Bowles 2010)
[39] (Fehr and Gächter 2002)

the game several times, but they told participants that they would never interact with anyone more than once.

The structure of this game creates incentives and costs that militate against donating much to the project. For instance, the best possible outcome for any one individual would be to invest none in a case where everyone else went all in. In that case, the free-rider would walk away with 44 MUs, whereas everyone else would gain a modest 4MUs, walking away with 24. Moreover, if I am the only one to contribute all my endowment, then I will end up with less than half my endowment, while everyone else turns a profit. This constrains how people actually play the game. In one experimental condition, after six iterations of this version of the game, Fehr and Gächter report, "...58.9% of the subjects contributed nothing and 75.6% contributed 5 MUs or less."[40]

In another version of the game, Fehr and Gächter introduced the possibility of punishment. They told participants that at the end of the round (after receiving information about how much others in the group invested) each participant had the opportunity to spend some of their MUs to punish another group member. For each MU contributed toward punishing an individual, that individual would lose three MUs (with the loss was capped at 30). With the possibility of punishment in place, the average investment immediately shot up to more than 12 MUs. On the sixth iteration of the punishment condition, Fehr and Gächter report, "...38.9% of the subjects contributed their whole endowment and 77.8% contributed 15 MUs or more."[41] Not only was the threat of punishment effective in securing cooperation, punishment also occurred quite frequently, with more than 80% of subjects punishing at least once across the six iterations of the punishment condition.

[40] (Fehr and Gächter 2002, 138)
[41] (Fehr and Gächter 2002, 138)

Conjecturing that emotions were responsible for this pattern of punishment, Fehr and Gächter followed up at the end of the game with questions about participants' feelings toward another player given how that player's investment compared to that of the rest of the group: "You decide to invest [5] francs to the project. The second group member invests [3] and the third [7] francs. Suppose the fourth member invests 2 francs to the project. You now accidentally meet this member. Please indicate your feeling towards this person."[42] Subjects rated both their anger and annoyance at the free-rider on a seven-point scale (one being "not at all" and seven being "very much"). Even with this rather modest discrepancy, 17.4% of participants indicated "very much" anger. Moreover, the higher the discrepancy between the investment of the free-rider and that of the other group members, the more anger participants reported (84% indicated five or greater in response to a vignette with a greater disparity between the free rider and others). Punishment in the public goods game followed a similar pattern. The higher the discrepancy was between the free-rider's investment and the average investment of the others, the greater the punishment (the more MUs that were lost). Moreover, the most common form of punishment in the game was when above-average investors punished below-average investors.

For my purposes, this study holds a pinch of evidence and a generous helping of illustration. It provides a pinch of evidence that anger is a reactive process that results in punishment and that punishment secures cooperation of the relevant kind. The study suggests that anger is a reactive process in that it leads to an impulse to punish in response to the *past* actions of a free rider even in cases *in which anticipated outcomes (such as maximizing profits) do not favor punishment*. This is plausible because participants knew that they would not encounter the free-rider again, yet participants punished free-riders proportionally to the severity of their free-riding. Moreover, they

[42] (Fehr and Gächter 2002, 139)

punished free-riders just as frequently on the last round of the game, in which no one in the game would benefit from punishment. Finally, the threat of punishment in the game secured the benefits of cooperation even between people who were unrelated, who had no idea of each other's reputation, and who had reason to think they would not interact with each other again. Importantly for my purposes, the study also illustrates how anger could possibly play an adaptive role in securing cooperation across a broad range of conditions *because it is a reactive process*. It supports a "how possibly" explanation for the evolution of altruism according to which a reactive processes was selected for its role in securing cooperation, and that is all I really need if I want to show how evolutionary considerations could, in principle, undercut deontological intuitions without undercutting evaluative intuitions more broadly.

2.3 An undercutting defeater for deontological intuitions

Let us suppose that this adaptationist story is correct and that anger is a reactive process. In order to secure the non-immediate consequences of cooperation, the reactivity of anger leads us to act based on the past action of the free-rider rather than because of immediate outcomes relevant to punishment (e.g. improved investment in a cooperative venture). In so acting, we manifest an inclination toward punishment that is not motivated by the consequences of punishment. If so, then in the context of punishment, anger includes a set of inclinations to act in accordance with R (to act as if the value of punishing did not depend on its consequences). However, if we are inclined to judge and act *as if* the value of punishment is not based on outcomes *precisely because* this feeling was adaptive for securing good outcomes, then the feeling is not trustworthy regarding R. There is a disconnect between, on the one hand, the value of punishment reported by intuition and, on the other hand, the manner in which these intuitions arose. According to R, punishment is supposed to have value apart from its consequences but these intuitions arose because of their consequences.

Suppose for the sake of argument that there is some state of affairs that makes punishment valuable independently of its consequences (or a state of affairs that makes R true). The retributive intuitions produced by anger could not possibly have an evidential connection to that state of affairs. Anger produces intuitions that support R because such intuitions deter freeriding, but the fact that these intuitions deter freeriding lacks an evidential connection to the states of affairs that R reports. More specifically, (and for reasons that I discuss in more detail in the following section) deterrence (or any other consequence of punishment) cannot be an *indicator* of any value that a punishment might have aside from its consequences (for reasons that I discuss in more detail in section 2.4). Thus, the putative evolutionary function of anger in the production of retributive intuitions serves as an undercutting defeater for those intuitions with respect to R.

Perhaps an analogy will flesh out the line of reasoning. Suppose that Geppetto is designing the psychology of a cyborg that he calls Pinocchio. Geppetto wants to make Pinocchio very realistic, and his aesthetic sensibilities favor a slightly scrawny boy. He foresees that this design preference will result in real boys picking on Pinocchio. Thus, he programs into Pinocchio a strong drive to resist bullies. He reasons that the policy of resisting bullies, even in cases where *immediate consequences militate against doing so,* will lead Pinocchio to suffer less from bullies in the long run. Bullies will realize that it is less costly to pick on other scrawny boys who are less scrappy, and they will bother Pinocchio less as a result. Geppetto wants Pinocchio to have the capacity for prospection, but Geppetto cannot guarantee that Pinocchio will be able to consistently anticipate the long-term value of resisting bullies.[43] Therefore, Geppetto designs Pinocchio with a drive to resist bullies that is not derived from the immediate prospective value of doing so.

[43] In any single encounter with a bully, Pinocchio would anticipate suffering immediate losses that he would not suffer if he did not resist; losses that favor giving in over resisting.

This drive gives Pinocchio an *urge to react* to the provocations of bullies rather than only *to respond to the immediate prospects (largely negative) of doing so*.[44] To Pinocchio, the urge to resist is there whether or not it will result in a good outcome, thus to him, the urge does not to derive from the anticipated outcome of resisting.[45] Once Geppetto completes his design, Pinocchio will be inclined to act and judge in accordance with the principle that resisting bullies has value not derived from its consequences or *non-derivative value*.[46] He might even discover that his intuitions about resisting bullies support the following principle and come to consciously believe it.

B - The value of acts of resistance toward bullies is not (or not only) derived from their consequences.

From the third-person perspective, it seems obvious that Geppetto's design has distorted Pinocchio's axiological beliefs about resisting bullies. If someone were to tell Pinocchio of Geppetto's design choices, he should no longer believe B. With the right information, he should conclude that his inclinations to resist bullies are not good evidence for B. Since Pinocchio's inclinations to resist bullies are disconnected from any source of value that resistance might have aside from its consequences, he has an

[44] Notice that when a desire is characterized as non-derivative in this way, it need not be indefeasible by consequentialist considerations. That is, overturning or defeating such a desire with a competing desire to maximize consequences does not make the defeated desire any more derivative. For instance, if a bully threatens lethal force, Pinocchio might overcome his urge to resist because of the catastrophic consequences of doing so. However, notice that this would not mean that the urge is derived from anticipated outcomes. That is, when Pinocchio has this urge prior to the threat of lethal force, it is not an urge to bring about an outcome. Rather, the right way to describe the situation is that Pinocchio feels an urge to resist that does not derive from the consequences of doing so, but he does not give in to that urge because of the catastrophic consequences of doing so. Retributive intuitions are similarly defeasible by consequentialist considerations. For instance, I suspect that most retributivists would say that even if punishment were good in itself, it would be reasonable and right not to punish someone because doing so would have catastrophic consequences. That is, retributive intuitions seem defeasible in just the way Pinocchio's non-consequentialist urge could be.

[45] Essentially, Geppetto programs Pinocchio with a subjective commitment device, a concept owing to the work of several economists, (e.g. Frank 1988).

[46] It is tempting to think that non-derivative value is identical to non-instrumental value. However, I think these concepts are distinct. For instance, a personal insult can be understood as instrumental for "getting even", but this is not to say that the value of the insult is derived from its consequences. Rather a successful insult is constitutive of "getting even". To me, this looks like an example in which an insult has non-derivative value (from the perspective of the insulter), but in which it is understood as instrumental for another aim, getting even.

undercutting defeater for those intuitions. Therefore, Pinocchio should not believe B on the basis of his intuitions.

If this argument is compelling in Pinocchio's case, then it should also be compelling in the case of R. If the adaptationist story about anger is correct, then the intuitions that support R secure good biological consequences in the long run just as they would if they were designed to do so. As such, they are similarly disconnected from any non-derivative value that punishment might have.[47] Thus, insofar as anger influences our intuitions about punishment, we are not justified in believing based on intuition that punishment has non-derivative value.

2.4 A more detailed explanation of the disconnect

The defeater I have offered is supposed to sever the evidential connection between retributive intuitions and the retributive principle R. How does this work? The argument shows that retributive intuitions are not a good indicator of non-derivative value (reported by R). To demonstrate this more clearly, let us take a closer look at the concept of indication. One state of affairs can indicate another if the states are highly correlated with one another, either *because one state of affairs causes (or constitutes) the other or because both states of affairs have a common cause (or constitutive base)* (Dretske 1999).[48] The requirement of a causal or constitutive dependency relation between two variables, is intended to rule out coincidental correlations between them. For instance, from 1999 to 2009 there was a strong correlation between the number of people who drowned in swimming pools and the number of films that Nicolas Cage appeared in (see www.tylervigen.com). However, this does not mean that Cage

[47] Notice that the argument does not hinge on any conflation of biological and moral goods. If the intuitions were shaped to bring about good consequences of any kind (e.g. biological or moral), then they cannot indicate non-derivative value of any kind, whether moral or biological.

[48] Dretske cashes out the dependency relation in terms of causation. I suspect that the dependency relation that explains a correlation need not be causal; it could also be constitutive dependency. See Berker forthcoming for a detailed discussion of the constitutive grounding relation in connection with debunking arguments.

appearances are an indicator of drowning deaths (or vice versa), because the correlation could be entirely coincidental.

Now, even if there were a correlation between retributive intuitions (the inclinations that support R) and non-derivative value (for punishment), the correlation could not possibly be explained by any of the relevant causal or constitutive dependency relations. Thus, any correlation will be coincidental and retributive intuitions cannot be an indicator of non-derivative value.

Consider the three possible explanations for the correlation. Suppose that retributive intuitions cause or constitute a state of affairs in which punishment has non-derivative value. If non-derivative value of punishment was constituted by the existence of retributive intuitions and if retributive intuitions were selected for their deterrent value, then the non-derivative value of punishment would depend on (either causally or constitutively) the deterrent value of retributive intuitions. Such a dependency seems impossible. Given the definition of non-derivative value, there should be cases in which punishment has non-derivative value but in which the deterrent value of retributive intuitions fails to obtain. For instance, there are possible contexts in which retributive inclinations do not deter freeriders. When those with retributive inclinations represent only a small percentage of a population, punishment for freeriding becomes so unlikely as to obliterate the deterrent effect of these inclinations (see e.g. S. Bowles and Gintis 2004). It seems that if punishment has non-derivative value, then it should retain its value even in those circumstances. But if this is correct, then any correlation between the deterrent value of the intuition and the non-derivative value of punishment will be a coincidence, due to the fact that most of the time, acts of punishment *just happen* to occur in a context in which retributive intuitions do have deterrent value.

Now suppose that the non-derivative value of punishment causes or constitutes the states of affairs in which retributive intuitions exist or are manifested. This

possibility seems to be almost entirely ruled out by the evolutionary explanation of retributive intuitions. If that explanation is correct, then we would have retributive intuitions whether or not punishment has non-derivative value. So it also seems unlikely that the manifestation of retributive intuitions depends in any way on the non-derivative value of punishment. If these intuitions were selected for their good outcomes, then there is no reason to think that their manifestation would depend on punishment having non-derivative value. Rather, their manifestation will depend on whatever conditions are necessary for them to have deterrent value. Of course, retributive intuitions might happen to be manifested in cases in which punishment has non-derivative value, but this would be entirely coincidental.

Now suppose that retributive intuitions (either their existence or manifestation) and the non-derivative value of punishment have a common cause.[49] This too seems impossible if the evolutionary explanation is correct. Again, the existence of retributive intuitions depends only on their deterrent effect. So it is hard to see how some other factor could cause retributive intuitions to exist and also cause punishment to have non-derivative value. Likewise, it is hard to see how a third factor could cause instances of punishment to have non-derivative value and cause retributive intuitions to be manifested. If these intuitions were selected for their deterrent function and if this function depends on the frequent manifestation of retributive intuitions, then it appears unlikely that non-derivative value would be caused by any of the conditions that elicit retributive intuitions. If the two happen to co-occur then this would be entirely coincidental.

The problem with all of these possibilities is that non-derivative value and deterrent value are, by their definition, independent sources of value. As a result, they

[49] We need not consider the possibility that they have a common constitutive base, because as a matter of definition, they do not.

are constituted by different facts, and any connection between the two sources of value will be coincidental. Non-derivative value is constituted by facts about an action aside from its consequences, whereas deterrent value is constituted by facts about the action's outcome aside from the intrinsic features of an action or what came before the action. Any overlap between these sources of value is likely to be coincidental.

2.5 What the argument does not show

Now, the argument might tempt someone to conclude that retributive inclinations are untrustworthy in the sweeping sense that we are never justified in acting or judging on their basis, but this conclusion certainly does not follow. A key feature of the argument is that it only severs the evidential connection between retributive intuitions and the retributive principle R. It does not undercut their evidential support of other beliefs or principles. The Pinocchio case also helps to illustrate this. Even if he should not believe that resisting bullies has non-derivative value, he may be warranted in acting on his urge to resist. Geppetto's forethought and beneficent design may give Pinocchio some warrant for acting on his urge to resist bullies. Likewise, the argument I have given gives us no additional reason to mistrust our retributive intuitions when applied to *practical* questions about when to punish (as opposed to *axiological* questions about whether it has non-derivative value). However, we cannot guarantee that natural selection positions us to have retributive intuitions with moral worth.

Here is another way to put the point. The argument debunks retributive intuitions as evidence for retributive standards of rightness, but does not debunk the intuitions as evidence for a retributive decision procedure. A retributive standard of rightness might say that punishment is right if and only if it is deserved (given the nature of a past offense). If the debunking argument is correct, then retributive intuitions do not provide good evidence for such a standard (at least insofar as this standard implies that punishment has non-derivative value). Nevertheless, this is not to say that retributive

intuitions do not provide good evidence for a retributive decision procedure, especially if such a procedure can be justified because of its good consequences (e.g. deterrence).

By analogy, Pinocchio may very well be justified in accepting the following decision procedure: if you have an urge to resist the bully, then do so. Such a decision procedure would be warranted insofar as Pinocchio is not a masochist and insofar as Geppetto's design tends to diminish bullying in the long run. Similarly, we may at times be justified in punishing on the basis of retributive intuitions (in accordance with a retributive decision procedure). This is because the consequences of doing so may align with our moral or non-moral aims. On either way of putting the point, the undercutting defeater only applies to the support that our intuitions seem to provide for R. I cannot think of any reason to think that they are untrustworthy in the more sweeping sense.

2.6 Generalizing the argument

Importantly, this conclusion only pertains to retributive moral intuitions, not to all deontological intuitions. Nevertheless, if the evolution of other moral emotions follows this same pattern (placing non-derivative value on action in order to improve fitness), then a similar argument can be given concerning deontological intuitions more broadly. Greene provides reasons to generalize.[50] The idea is that moral emotions are domain-specific adaptations, where the specific domain of each moral emotion is a recurring situation that constitutes one of the "…demands and opportunities created by social life."[51]. In programming us with emotions, nature declines to "…leave it to our powers of reasoning to figure out that saving a drowning child is a good thing to do…"[52] or that hurting others and lying are bad things to do. In other words, many moral emotions lead us to *react* to specific kinds of situations (ones that involve *inter alia* assistance, punishment, promises, testimony, and incentives to harm) in specific ways

[50] (See esp. J. D. Greene 2008, 59–60, 72)
[51] (J. D. Greene 2008, 60)
[52] (J. D. Greene 2008, 60)

rather than responding only to the prospective value of acting. This is because the immediate prospects these situations present lean in favor of declining assistance, avoiding confrontation and punishment, breaking promises, lying to others, and doing physical harm to get what one wants. Moreover, acting against these inclinations is supposed to be adaptive for its role in supporting human cooperation. If this is right, then we can generalize the argument to undercut intuitions that support a broader range of deontological principles, specifically, any deontological principle that the value of a certain action is not (or not only) derived from its consequences. The case of anger and punishment allows us to see the kind of problem that would be raised for other deontological intuitions. Nevertheless, from a methodological perspective, the case needs to be made one moral emotion and one deontological principle at a time.[53]

3. Objections and replies

3.1 The argument proves too much

The most obvious objection to the argument is that there is also a disconnect between evaluative intuitions and the sources of value they report. For instance, nociceptive processes produce intuitions that seem to support the claim that that pain is bad, and they were selected to do so because of their tendency to aid survival. Here, we may have a similar disconnect. The intuitions report the *objective* badness of pain, whereas nociceptive processes produce those intuitions because of the *biological* badness of bodily insult and injury. If there is a disconnect between objective badness and biological badness, then we have an undercutting defeater for these evaluative intuitions. Therefore, the objection goes, the case against non-prospective processes seems to again prove too much, since it also undercuts the evaluative intuitions that

[53] For one, I am not sure that the moral emotions can all be explained in this way. For another, it is not yet clear to me that all cases of non-derivative valuing derive from domain-specific adaptations of this kind.

support a theory of value (necessary for consequentialism). In other words, it does not undermine deontology any more than it undermines consequentialism.

3.2 Reply: the putative disconnect between evaluative intuitions and objective values is different

While evaluative intuitions may be on shaky footing, their footing is independent of deontological intuitions. This is because my argument against deontological intuitions points out a distinct undercutting defeater. The undercutting defeater severs the evidential connection between deontological intuitions and non-derivative value by showing that deontological intuitions cannot indicate non-derivative value. The defeater severs this specific evidential connection *not only* because deontological intuitions were shaped by evolution but also because of the content of the deontological principles in question: that certain actions have non-derivative value. If an inclination was selected for good outcomes like deterrence, then it cannot indicate that acting in accordance with that inclination has non-derivative value (as deontological principles state).[54] Thus, I am giving a reason to doubt the evidential value of deontological intuitions (with respect to deontological principles) that we do not have for doubting other evaluative intuitions (which support other principles besides deontological principles).

It is an entirely different question whether an inclination that produces good biological outcomes can indicate that certain actions have moral value (regardless of whether this value is non-derivative). This is one question that the evolutionary debunking arguments of Street attempts to answer. Moreover, it is a question that does not hinge on whether non-derivative value enters into the content of the evaluative

[54] At least, this is true so long as non-derivative value is understood directly in terms of the value of acting and not in terms of value that is derived from the acceptance of a certain ethical decision procedure or practice (which might have a consequentialist justification). Consider an example. Rawls (1955) distinguishes between justifying actions *within a practice* of promising or punishing (both of which involve ignoring certain reasons one might have not to promise or not to punish) and justifying the practice itself (perhaps in terms of its consequences). Reasons that one accepts from within a practice (and which are in some sense supported by the consequences or aims of the practice) do not count as non-derivative reasons if they are grounded in the consequences of the practice.

beliefs in question. Moreover, my argument focuses on the relation between derivative and non-derivative value, whereas these debunking arguments focus on the relation between biological and moral value. There are important differences between these relations. As I argued above (in section 2.4), non-derivative value and deterrent value are, by their definition, independent sources of value. By contrast, there is room in conceptual space for biological goods like survival to (partially) constitute objective goods like flourishing.

The point is that the defeater that I have proposed for deontological intuitions severs their evidential connection with deontological principles in a way that does not apply to evaluative intuitions more broadly, so the debunking argument does not prove too much. Thus, we have a clear case in which evolutionary considerations have implications for first order moral discourse, somewhat independently of metaethical concerns raised by the evolution of our moral faculties.

4. Conclusion

In their efforts to undercut some of the primary evidence against their theories, some consequentialists have employed evolutionary debunking arguments against deontological intuitions. So far, these arguments have met with little success, because the evolutionary considerations offered may undercut a much wider range of intuitions. I presented an argument that overcomes these challenges. I argued that non-prospective processes – processes that motivate action but not based on an action's outcome – might explain a range of deontological moral intuitions. These intuitions seem to support anti-consequentialist principles, according to which the value of various kinds of action does not derive from that action's outcome. Nevertheless, the evolutionary function of some non-prospective processes is to bring about certain outcomes. If this is right, then it is quite plausible that we would have these intuitions even if they were false, even if the relevant actions were only valuable because of their consequences.

Of central importance for the argument I have just given, the defect in non-prospective processes is not just that they aim at reproductive fitness, but more specifically that they increase reproductive fitness *by influencing organisms to react to certain types of situation or action* rather than approach them prospectively. The adaptiveness of this tendency reveals a disconnect between the intuitions that these processes produce and the states of affairs they seem to report (through their support of principles like R). So far, this specific defect only undermines deontological intuitions. The argument does not seem to apply to evaluative intuitions more broadly because biological values and objective sources of value do not necessarily have an independent constitutive base.

Chapter 2
Spiteful Punishment and Retribution: Strategy and Motive for Non-Strategic Organisms

Kathi Sylvester was working as a crossing guard one day at a school in Surprise, Arizona when Timothy Francisco approached the crossing in his car. He signaled no intention to stop, and in response, Sylvester "smacked" his driver-side window and brandished her stop sign, thereby urging him to obey the traffic regulations for school zones. In reaction, Francisco got out of his car and punched Sylvester in the face, leading to a broken jaw and an overnight hospital stay.

One interesting feature of Francisco's actions (and perhaps Sylvester's as well) is that they were clearly reactive. In response to her unintended provocation, Francisco imposed a serious cost on Sylvester in the form of physical and emotional harm, and it is unlikely that he acted because of the anticipated outcome of his action. Even if Francisco thought he could get away with assaulting Sylvester, it certainly wouldn't have done him any good to do so. If he was in a hurry to get somewhere, as the details of the case suggest, stopping his car and getting out probably put him in a bigger rush when left the scene of the assault. If he was concerned about his reputation, he certainly would have gained little by assaulting someone who had not made any serious threats to it. Even if the assault had reputational effects, it is unlikely that as a result of his augmented reputation he would deter enough poor treatment to offset the probable costs of assaulting Sylvester. Additionally, the more widespread his reputation became for this incident, the more likely that Francisco's crime would be detected by authorities. While it is not entirely unlikely that Francisco got pleasure or satisfaction from punching Sylvester, neither would this seem to balance out the risks of assaulting her.

Of course, this could all derive from short-sightedness. On this hypothesis Francisco acted with some of the outcomes of his action in mind, but did not think through all of the possible outcomes. When Francisco later turned himself in to the

authorities, he admitted that he let Sylvester "get to him". This amounts to an admission that he should not have let her "get to him" and that he would not have acted so impulsively if he had not. By his own admission, Francisco's actions reveal a regrettable lack of foresighted control. Additionally, the salience of institutionalized punishment for crimes like assault makes it unlikely that Francisco thought through the possible costs of his action. We need an explanation of why these salient costs were obscured or outweighed by the desire to retaliate. Why did the impulse to react to provocation overwhelm Francisco's foresight? A salient possibility is that Francisco was angry and that his anger subverted the more careful prospective deliberation that he might have otherwise engaged in.

Francisco is not alone in the possession of angry, reactive impulses, and this is part of the reason that blame seems an appropriate response to his impulsive action. At times, anger impels each of us to react against our better judgment if we do not actively resist the impulse. Why is this so?

Many suspect that biological or cultural evolution offer the best explanation for reactive, retributive impulses connected with anger (see e.g. Frank 1988; Petersen et al. 2010; McCullough, Kurzban, and Tabak 2012). This is the sort of explanation appealed to in the how-possibly explanation of retributive intuitions in the first chapter. One peculiarity of the view presented there is that the retributive motivation inherent to anger was presented as a group level cooperative adaptation, one that helps to protect cooperation by motivating retributive punishment of defectors, or punishment that is not motivated by the anticipated outcomes of punishment. Nevertheless, as we saw in Francisco's case, the reactivity of anger is not secluded to its influence on moral punishment – directed at norm violations. It is unlikely that it only functions to enforce the kind of extreme cooperation (in non-iterative, non-reputational interactions with non-kin) that we see in human beings. Rather, a psychological state like anger may also

have a role to play in the tendency of a vervet monkey to attack the kin of the monkey who just attacked him (e.g. Cheney and Seyfarth 1989); or in a capuchin monkey's rejection of unequal rewards (e.g. Brosnan and de Waal 2003); or perhaps even in the behaviors of more distant mammalian lineages. In this chapter, I address this peculiarity in the evolutionary explanation of the previous chapter. I argue that current evolutionary explanations of retributive punishment are actually insufficient to explain why anger includes a retributive motive and why this motive produces an impulse that easily results in spiteful action. I then present a novel proposal concerning the origin of this disposition.

In the following section, I point out how current evolutionary explanations of retributive motives rely on the prospective capacities of organisms. On these models, punishment is understood as a behavior modification strategy, and the prospective capacities of the punishee are necessary for the strategy to succeed. I pose two problems that arise from this understanding of retributive punishment, and I argue that these problems point to the necessity of a more ancient evolutionary explanation of the retributive motivation inherent to anger, one that precedes the development of complex prospective capacities. In the second section, I develop just such an explanation. This explanation demonstrates the necessity of a retributive motive to implement a uniquely stable strategy for resource competition. In the penultimate section, I show that this model does explain certain features of an anger-like motivational state of rodents. Nevertheless, I will not make the argument that this motivational state is of the same kind as anger in humans, a task I leave for the following chapter. I conclude by abstracting from this case some of the benefits of understanding the phylogenetically ancient adaptations that humans may possess.

1. Limitations on Idealized Models of the Evolution of Punishment

In this section, I argue that prominent models explaining retributive motivation suffer from explanatory deficiencies. To make that argument, I need to say a good deal more about these models. The models can be used to explain, among other things, the existence of *punishment*.[1] I understand punishment as a behavioral strategy by which one entity (e.g. organism or group of organisms) diminishes the fitness of another entity in response to (or proximally caused by) a provoking incident involving that entity, usually an incident that diminishes the fitness of the punisher (cf. Nakao and Machery 2012). I leave the details open to broad interpretation, because I think there are several possible types of punishment that are conceptual possibilities. For instance, the definition is satisfied by one organism diminishing the absolute, relative or inclusive fitness of another organism at some or no cost and is even satisfied by a group diminishing the fitness of another group or a group diminishing the fitness of an individual. Behaviors like these all seem to me to count as punishment. Nevertheless, the definition does exclude predation behaviors. While predators do diminish the fitness of their prey, they do not usually do so in response to a provoking incident. The models I will consider have two other features that require further elaboration. They require prospective capacities on the part of the punishee, and the punishment strategies they explain cannot be implemented by prospective capacities.

1.1 Prospective capacities of the punishees

The models of human punishment that I will criticize explain its existence and maintenance as a behavior-modification strategy. This kind of explanation is pursued by a majority of evolutionary models of punishment (see e.g. Nakao and Machery 2012) with the exception of those I consider shortly. The most plausible proximate explanation for the ability to learn from punishment (on the part of the recipient or audience of

[1] The phenomena on which many of these models focus is actually human cooperation. Nevertheless, on these models, punishment in some form seems to be necessary for certain forms of cooperation and cultural evolution. As such, the models can be used to explain the existence of punishment as an adaptation for cooperation.

punishment) is the possession of prospective capacities. Recall that prospective processes select actions based on the internally represented outcome of the action. Moreover, the consequences of actions are calculated with reference to a learned causal model (e.g. Gläscher et al. 2010). The selection models that I criticize require that agents be able to learn causal relationships to anticipate the outcome or expected utility of social interactions. When applied to these models this means that a punishee or audience can acquire information from the punisher, from the act of punishment or from its consequences and can use that information to tailor some of their decisions to the strategic demands of each situation. For example, self-interested agents with prospective capacities can learn from punishment of themselves and others that certain selfish behaviors leads to punishment at least when a known punisher is present. As a result, the likelihood of punishment can offset the material benefits of selfish behavior in the relevant situations and in a way that coerces self-interested agents to modify their behavior.

Importantly, successful punishment strategies in models of direct reciprocal altruism (as in Axelrod 1984; Trivers 1971) often do not require behavior modification, because they involve forms of punishment either that benefit the punisher by cutting their losses or that affect interaction partners without modifying their behavior.[2] Thus, the problem I will pose does not apply to many of the models within this broader family.

On the other hand, reputational models usually do require agents with prospection. Consider the commitment model (Frank 1988). This model attempts to explain how humans are able to solve commitment problems like the following:

> **Deterrence.** Suppose Smith grows wheat and Jones raises cattle on adjacent plots of land. Jones is liable for whatever damage his steers do to Smith's wheat. He can

[2] For example, an organism can punish another organism by killing it. For a range of biological examples of this form of punishment, see Nakao and Machery (2012)

> prevent damage altogether by fencing his land, which would cost him $200. If he leaves his land unfenced, his steers will eat $1000 worth of wheat. Jones knows, however, that if his steers do eat Smith's wheat, it will cost Smith $2000 to take him to court…Smith threatens to sue Jones for damages if he does not fence his land. But if Jones believed Smith to be a rational, self-interested person, this threat is not credible. Once the wheat has been eaten, there is no longer any use for Smith to go to court. He would lose more than he recovered. (Frank 1988, 48)

The problem concerns how Smith can make a credible threat, when Jones knows that it is not in Smith's immediate interest to go to court. In a case like this, a retributive motivation can serve as a commitment device[3], meaning that it would motivate Smith to follow through with his threat and thereby punish Jones, even if it is not in his immediate interest to do so. Smith is likely to benefit in one of two ways from possessing this commitment device. One possibility is that Smith will retaliate against Jones's noncompliance and thereby foster a beneficial reputation for following through with his threats. (For instance, if Smith's neighbors Amjad, Baggi, and Castro each subsequently decide to raise cattle on their property, then Smith's threats will be more credible to them than they were to Jones, providing a net benefit to Smith if they comply with his wishes.) The other possibility is that Jones complies because, as a result of Smith's retributive motivation, he already has such a reputation.[4] Either way, Smith is likely to get reputational benefits from retributive motivation.[5] Moreover, it is easy to see how

[3] More specifically, it is a subjective commitment device, contrasted with objective or external commitment devices, which might offer, say, material incentives for following through with a threat. For instance, Smith might put the money necessary to sue Jones in an escrow account for his lawyer, the reimbursement of which is contingent on Jones building a fence. In this way, Jones would "tie himself to the mast" with an external commitment device.

[4] Frank (1988) also goes to great pains to show how reliable signals of moral sentiments like outrage might be achieved, so that those who possess them rarely violate their self-interest despite having a tendency to do so.

[5] One might expect that the reputational benefits of punishment would be factored into the decision to punish and that this would make the decision purely strategic in the sense of being guided by internally represented future benefits. However, punishment motives need not be prospective in order to be sensitive

these benefits could translate to fitness benefits. This kind of explanation of retributive motivation requires prospection because the benefits of punishment only obtain if other agents are influenced by the reputation of the punisher and if they can use this information to tailor some of their behaviors to the demands of situations involving that specific individual. On this model, the reputational benefits have their effects by entering into the strategic considerations of those deciding whether to comply when threatened.

Also among the models I will criticize are prominent group selection models, which aim to explain the evolution of human cooperation by appealing to the group level benefits of temporarily spiteful punishment (e.g. Robert Boyd, Gintis, and Bowles 2010; S. Bowles and Gintis 2011). These models generally focus on large, cultural groups and demonstrate that a strategy of retributive punishment can protect cooperation in large groups or prevent group extinction and thus enable cooperation to persist and spread as a result of cultural group selection. On these models, retributive punishment ends up being adaptive because it correlates with group cooperation, which can yield higher fitness benefits to punishers than the relative within-group costs of punishing. The latter would be very low when most people cooperate or when there are low cost forms of punishment such as social gossip and ostracism.

These models depend on prospection because the models evaluate the conditions in which purely self-interested agents will cooperate due to the *likelihood* of punishment. This assumption requires prospection because the evolution of punishment in these

---

to factors that are strategically relevant to the effectiveness of the commitment device. For instance, audience effects, whereby a behavior is mediated by the presence of other individuals, typically can be explained by the eliciting conditions of the behavior rather than by an anticipation of the effects on the audience. Many non-human species display audience effects (e.g. Marler, Dufty, and Pickert 1986) but nonetheless lack the ability to represent the mental states of other organisms. Moreover, there is reason to suspect that some human emotions are unconsciously influenced by the presence of an audience (Fridlund 1991). Thus, there is good reason to believe that reputational effects can be achieved without the significant computational costs of representing their effects in prospective deliberation. Thus, there is good reason to suppose that punishment motives can be sensitive to strategically relevant factors without any internal representation of the strategic effects of those factors.

models depends on individuals being able to predict the likelihood of punishment given prior interactions with or observations of punishers.[6]

My criticisms will also apply to recent partner selection models. These models explain the evolution of cooperation by appealing to partner selection as an evolutionary pressure resulting from the necessity of engaging in mutually beneficial cooperative ventures (e.g. Baumard, André, and Sperber 2013). In these models, individuals succeed by being selected as partners in mutually beneficial ventures. Whether an individual selects a potential partner depends on whether the individual appears attractive as a partner. For instance, if someone is known to be untrustworthy in various ways, he will be less attractive as a partner because other individuals know that the he may defect or unfairly divide or conceal the returns of the mutual venture. By contrast, individuals with a tendency to punish unfair behaviors would be highly attractive accessories in mutually beneficial ventures. Prospective partners know, for instance, that they can trust such a person to oppose or punish someone else who attempts to make off with an unfair share of the returns. Given that attractive partners have increased likelihood of success and thus survival, we can explain the widespread tendency to punish unfair behaviors in terms of this tendency. Obviously, partner selection requires some degree of prospection in order to decide which partner is the most attractive for a collaborative venture. The partner must be able to anticipate the likelihood that potential partners will act in certain ways in order to decide which partnership will produce the most profitable outcome.

1.2 Punishment is not motivated by prospection

While all of these models require prospective capacities of punishees, they indirectly constrain the use of prospective capacities *in punishers*. In all of the models that I will consider, punishment has a cost and the benefits of punishment, if any, are

[6] Moreover, the explicit target of these models are groups of hunter-gatherers in the Pleistocene, well after the human and chimpanzee lineages had diverged and during which tool use became more common (e.g. Samuel Bowles, Choi, and Hopfensitz 2003). Thus, it is also likely that the organisms to whom these models apply had prospective capacities.

deferred or intangible. This means that punishment in these models is often temporarily spiteful, because it imposes a cost on the punishee at an immediate cost to the punisher. Importantly, the punishment strategies that are successful on these models cannot be implemented by the prospective capacities that we find in human and non-human animals. Prospective processes will usually weigh against temporarily spiteful punishment because of its immediate material costs and lack of immediate material benefits.

One might think that prospective capacities could implement temporarily spiteful punishment so long as an organism is capable of long-distance prospection or extreme foresight. For instance, one might prospectively balance the immediate costs of punishment against its future consequences, say deterrence, and decide in favor of punishment[7] Nevertheless, well-established, cross-species work on temporal discounting shows that the prospective capacities of almost any species place exponentially greater weight on immediate rewards than more distant rewards (e.g. Critchfield and Kollins 2001). For example, the far off, intangible consequence of deterrence will as a matter of psychological fact, usually have much less value in the calculation of expected utility than the immediate costs of punishing. Moreover, for punishment to proliferate in the models in question, it has to take place in precisely this kind of situation, in which its benefits are far off or intangible. Thus, long-range prospection as it is implemented in most animals could not plausibly function as the proximate motive for temporarily spiteful punishment that these models aim to explain.[8]

Alternatively, one might think that punishment could be motivated by the prospective value of satisfying retributive preferences, such as the preference that non-

[7] It is important to notice here that prospection is not the same as foresight. Calculating the *immediate* benefits of an action by itself can be a prospective enterprise. For instance, one could use a causal model to calculate the immediate rewards of one action compared to another without giving any thought to more temporally distant rewards. This doesn't require a great deal of foresight, but it is a clear case of prospection on my understanding of that term.

[8] For a more detailed discussion of these issues, see Frank's discussion of the matching law (1988, ch. 4).

cooperators be penalized for their non-cooperation. The proposal is that this preference could offer its own non-material incentive that weighs against the material incentives not to punish. There are two problems with this proposal. One is a consequence of the fact that this is a non-material benefit. It follows that the reward function for this preference is not determined by any material benefit, like food or sexual fulfillment. That is, the aim, achievement of which satisfies the retributive preference, cannot be described in terms of future material benefits. Moreover, we cannot describe the object of these preferences without referring to the past: righting a past wrong, settling a past score, meting out just deserts of past actions, paying back past insults, balancing the scales that were upset in the past. While it is reasonable to think that the future satisfaction of these preferences could factor into the operation of prospection, the nature of these motives is still essentially reactive. Their satisfaction depends on reacting to past events rather than achieving an aim that refers entirely to the future. Thus, there is an ineliminable reactive component to this "prospective" implementation of temporarily spiteful punishment.

There is another problem with the idea that punishment is motivated by prospection that considers retributive preferences. It is that a retributive preference, as prospectively weighed against other incentives, is not a satisfactory explanation for many instances of temporarily spiteful punishment. For instance, in the case of Timothy Francisco, we have some reason to doubt that he engaged in prospective deliberation (or perhaps even subpersonal prospective computations) that weighed the satisfaction of his retributive preference against other concomitant outcomes. It is unlikely that the satisfaction of retributive preferences could have outweighed his other preferences (e.g. not to be imprisoned for assault). The bypassing of prospective considerations seems to be a general aspect of impulsive actions.

We can conclude that temporarily spiteful punishment is a behavioral strategy that probably requires a non-prospective motive, largely because temporarily spiteful

punishment is not directed at the immediate material benefits of punishment. In other words, temporarily spiteful punishment requires retributive motives for its implementation. Thus, evolutionary models that appeal to the adaptiveness of temporarily spiteful punishment strategies offer a promising way to explain the existence of retributive motives, since these motives are necessary to implement successful strategies.

Importantly, this claim does not apply to prominent models of reciprocal altruism (e.g. Axelrod 1984), because successful punishment strategies in these models usually do achieve an immediate material benefits and thus do not require a retributive motive. For instance, Axelrod's model of reciprocal altruism tested several different strategies in an idealized game called the prisoner's dilemma. In the prisoner's dilemma, players can cooperate or defect. If both cooperate, both receive a modest payout say $8. If one defects while the other cooperates, the defector receives the maximum payout say $10, whereas the cooperator receives the lowest payout, say $1. Finally, if both players defect then both receive a payout that is not the lowest but is less than the payout if both cooperate say $3. In the one-shot prisoners dilemma, the best strategy is to defect. Regardless of what the one player does, the other player will improve her payout if she defects. Suppose for instance, that my partner defects, if I cooperate, I get $1, and if I defect, I get $3. Now suppose that my partner cooperates, if I cooperate, I get $8, and if I defect, I get $10. In both cases, I do better if I defect. However, Axelrod found that when the dilemma is iterated with a specific partner, unconditional defection is not the optimal strategy. Specifically, a tit-for-tat strategy outperformed 40 other strategies carefully designed to defeat it. This strategy begins by cooperating, then replicates what its partner did on the previous iteration of the dilemma. This strategy does punish defection in one round by defecting in the subsequent round, but importantly, this punishment is not costly since the payoff for defection is better regardless of which choice the other

player makes. Moreover, this strategy is more successful than other strategies because it achieves good outcomes when the other player is cooperating and cuts losses otherwise. Rather than being temporarily spiteful, this punishment strategy benefits the punisher *and* allows the punisher to avoid losses.

The success of this kind of strategy is thus unlikely to explain retributive motives for punishment. For instance, Trivers (1971) does posit a retributive motive that functions to ensure reciprocation in situations like the prisoner's dilemma, but Frank (1988, 37–38) argues that this motive is more costly than a tit-for-tat strategy, which can be motivated purely by self-interest. A retributive motive would presumably lead to punishment of a non-cooperator even at a cost and regardless of whether one will interact with the individual again. If, on the other hand, one is motivated purely by self interest, one would only punish a defector to achieve a higher payout, to cut losses or coerce the partner to cooperate (as in tit-for-tat). Moreover, one would only punish at a cost when interactions are repeated, otherwise it would lead to immediate losses that would not be recouped by cooperation in subsequent interactions. The point is that tit-for-tat strategies do not require retributive motivation to serve their purpose in models of reciprocal altruism, so these models do not provide a satisfying explanation of retributive motives.

1.3 Problem 1 for models with prospection: presupposing the explanandum

The dependence of evolutionary models on prospective capacities creates two problems for models of agents with prospection. One is that in these models a great deal of the explanatory burden rests on the assumption that there is a range of strategies concerning punishment including a strategy of temporarily spiteful punishment. For any selection model, a certain amount of variation needs to exist for selection to operate on it. In this case, the models evaluate the success of different punishment strategies, and in doing so, they presuppose the existence of the range of strategies they evaluate. Of course,

it is a well-known limitation of selection based models that they assume a certain amount of variation in traits rather than explaining it. Moreover, the ability of a selection model to explain the target phenomenon depends on the plausibility of the relevant assumptions (Sober 2009). However, even if these assumptions are plausible for the selection models I consider, the existence of temporarily spiteful punishment as a variant requires a good deal of explanation in its own right. This is because of the complex capacities required to implement this strategy in organisms with prospective capacities.

Notably, the motive for temporarily spiteful punishment must override the prospective motivation to avoid immediate costs. While these models require organisms with prospective capacities to influence the punishee, the decision to punish cannot be motivated by the prospective capacities that humans and other animals possess. This is because the inclination to punish needs to persist despite the immediate losses that attend it. Given that prospective capacities benefit organisms in a very broad range of contexts, it is plausible to assume that prospective capacities will be applied to many different kinds of decisions, including the decision to punish. Insofar as organisms do apply their prospective capacities to decisions about punishment, they will be less likely to punish. Thus, to implement a strategy of temporarily spiteful punishment in organisms with prospective capacities, the influence of these capacities has to be mitigated in the domain of punishment. At the very least, we need an explanation for how organisms with prospective capacities would reliably punish at a cost when the benefits are deferred. It seems that the best explanation of this phenomenon in humans is that retributive motives override or influence prospective capacities.

This is problematic because a trait like this has a high degree of *interaction complexity* in that it requires a complex interaction with prospective capacities in order to override what material self-interest would demand in the context of decisions to punish. It is hard to see how this sort of trait could arise *de novo* (e.g. from a single point

mutation or from several simultaneous mutations) in an organism with prospective capacities. It seems unlikely that a small change in genetic material could lead to a complex tendency to punish (e.g. at a cost and only under certain conditions) *including* a complex interaction with prospective capacities (e.g. overriding or strongly influencing the decisions to punish).

Rather, it seems much more likely that in this case evolution had some prior material to work with from which to fashion this more complex trait. For instance, there is a widespread tendency in the animal kingdom to react aggressively when conditioned rewards are withheld (e.g. Looney and Cohen 1982)). Likewise, there is a widespread tendency to respond aggressively to provocation. For instance, when an animal is attacked by a conspecific, there is a tendency to engage in counter-aggression or to redirect aggression toward another conspecific (Barash and Lipton 2011). However, if this is correct, then our ancestors may have already had something quite like a retributive motive for natural selection to have acted upon prior to the development of complex prospective capacities. Yet if that is the case, then these explanations are not satisfying explanations of why humans have retributive motives in the first place. Instead they are explanations for why that trait was co-opted, and thereby maintained, to protect cooperation, to attract cooperative partners or to solve commitment problems. Thus, even if the models make plausible assumptions about ancestral variation in punishment strategies, they still may not offer a satisfying explanation for the *existence* of retributive motives, because it seems likely that such motives already existed prior to the contexts in which these other functions became adaptive.

1.4 Problem 2 for models with prospection: leakage

Another problem for these models was first introduced by Jon Elster:

> …when assessing the impact of an emotional disposition on reproductive fitness, one has to take account of *all* effects of the disposition, indirect as well as direct. In the

> case of anger, for instance, it may be true that irascible people often get their way, but that is only part of the story. Others will learn to recognize them as irascible and walk around them rather than have any dealings with them. Sometimes one has no choice, but often one can find alternative and more reasonable partners. Irascible people will find themselves shunned, which detracts from opportunities for mutually favorable interactions with others… I am not saying that the net effect of irascibility is negative, only that one cannot show it to be positive simply by citing a positive impact in isolation from other effects. (Jon Elster 1999, 48–49)

Here, Elster's primary target seems to be accounts like the commitment model. Yet he subsequently generalizes the point to selection based explanations and, in particular, their use of idealized models to explain emotions:

> …whether the account relies on the signaling function of the emotions, on their ability to underwrite the credibility of threats and promises, or on their efficacy in sustaining motivation over time, it would just be another story. Modeling always implies simplification…Yet when the question is whether a phenomenon exists because of its *net* reproductive benefits or in spite of its *net* reproductive costs, telling a story to demonstrate that it has *some* benefits or costs is not likely to be useful. (Jon Elster 1999, 49–50)

This problem applies to each of the models I discuss above. Each model considers a limited range of idealized situations involving punishment. In the commitment model, the focus is the problem of deterring via threat backed by reputation. In group selection and partner selection models, the focus is on a tightly constrained form of cooperative venture (with fixed cost and return functions). Moreover, the models evaluate the benefits and costs of punishment with respect to these idealized situations *in isolation from other kinds of situation.*

Simply put, the problem is one of leakage. We cannot guarantee that the best solution to a certain idealized problem situation will not have negative effects that leak into other types of situation. Yet if a trait has more net costs in *total* than a competing trait, then it will not survive natural selection regardless of whether it would be the most successful response to the one type of situation considered in isolation. Since these models do not account for the potential effects of punishment strategies in all situations in order to weigh all of the costs and benefits of retributive punishment, they leave us with an incomplete (though not necessarily inaccurate) explanation of why retribution survived natural selection.

Importantly, Elster's argument is most problematic for models of the evolution of organisms with prospection. The main example Elster gives to support the potential leakage from one type of situation to another is that individuals can learn to recognize irascible people and avoid interacting, presumably because of the likely outcomes of such interactions. That is, individuals can tailor their decisions to the strategic demands of each situation, treating situations differently based on the individual characteristics (e.g. irascibility) of the person with whom they might interact. I think the problem is this: when we introduce the ability to anticipate outcomes based on causal relationships (e.g. one that includes the information that irascibility increases the likelihood that the irascible person might treat me poorly), the effects of an agent's actions in one situation can have many more effects on the agent in other situations. For instance, suppose someone is trying to choose a mate and that Brown is known to be irascible and Mendez is not. As a result of his irascibility Brown might be less likely to be chosen than Mendez. In general, when you introduce the ability to anticipate an outcome based on a causal model, almost anything could become relevant to anything else.[9] It follows that an

[9] The frame problem in artificial intelligence and robotics depends on the fact that almost anything could be relevant to the outcome of an action. That is, the computational difficulties inherent in the design of robots

agent's behavior in one kind of situation could have effects on that agent (via the prospective abilities of her peers) in an indeterminate range of other situations.

By contrast, in organisms without complex prospective capacities, there is a more limited ability to learn from experience and to predict the consequences of behavior on the basis of experience.[10] As a result, there are fewer social interactions where an organism can make decisions based on the specific characteristics of its partner. If an organism cannot anticipate the consequences of getting on the wrong side of an irascible individual, then it will not be able to avoid that individual, and the consequences of irascibility will not as readily leak into other kinds of situation. This is why the kind of model I consider in the following section offers safer idealizations.

2. Safer Idealizations: Models of Resource Competition in Non-strategic Organisms

The models I consider in this section are ones that do not depend on the prospective capacities of organisms. In the organisms to which these models apply, there is less leakage between different types of social interaction because there are fewer ways in which an individual's behavior in one interaction can influence the outcome of other interactions and interaction types. Moreover, insofar as these models can explain the adaptiveness of a retributive motive prior to the evolution of complex prospective capacities, they mitigate the problem of interaction complexity posed above. Moreover, they make it possible to understand how a retributive motive could have co-evolved with prospective capacities. Thus, these models may also provide a basis for understanding how subsequent interaction complexity might have gradually arisen.

---

that can interact with the world in real time derive from the intractability of explicitly representing all of the factors that might be relevant to a candidate course of action. See e.g. Dennett (1984)

[10] While many species, including rats have capacities for prospection, the capacity to represent the value of an outcome is limited to the immediate future. Moreover, the contingency relations that animals are capable of learning are constrained by the temporal contiguity of the action and its effect (among other things). It should be clear that these limitations do not allow many nonhuman animals to learn from punishment in the manner required by the selection models discussed above.

In this section, I describe a model that can explain the necessity of retributive motives in non-strategic organisms. It is a game theoretical model that captures some of the dynamics of *frequency dependent selection*. This kind of selection exerts different pressure on a trait depending on the frequency of variant traits in a population. Biologists can evaluate the effects of this kind of selection on social interaction strategies using computer simulation of game theoretic models. Given various strategies for interacting in games like the prisoner's dilemma, one can use computer simulations to evaluate the average payoff of a strategy when played in populations consisting of organisms with various other strategies.

For instance, we could easily evaluate the relative success of a tit-for-tat strategy in a population consisting entirely of serial defectors. If we imagine that the prisoner's dilemma is repeated ten times for each encounter and there are ten encounters over an individual's lifetime, we need only sum up the outcome of each encounter with a serial defector and multiply it by ten. When tit-for-tat goes up against serial defection, it get $1 on the first round and the defector gets $10 because tit-for-tat starts by cooperating. For the other nine rounds, tit-for-tat will match the defector's behavior and both will get a payout of $3. So we get 1+3+3+3+3+3+3+3+3+3 x 10 = $280 for tit-for-tat as compared with 3+3+3+3+3+3+3+3+3+3 x 10 = $300 for serial defection (since the serial defector will only encounter other defectors in this population). Likewise, we could easily evaluate the success of tit-for-tat in a population half of which are serial defectors and half of which are tit-for-tat strategists. Given that the probability of interacting with a serial defector is 0.5, we just add up the expected number of encounters with serial defectors and add it to the expected number of encounters with other tit-for-tat strategists. We get (1+3+3+3+3+3+3+3+3+3) x 5 + (8+8+8+8+8+8+8+8+8+8) x 5 = $540 for the average tit-for-tat strategist as compared with (10+3+3+3+3+3+3+3+3+3) x 5 + (3+3+3+3+3+3+3+3+3+3) x 5 = $235 for the average serial defector.

In these models, successful strategies are more likely to be present in subsequent generations, and so these models can tell us which strategies selection is likely to favor given the frequency of other strategies in the population. The concept of an evolutionarily stable strategy (ESS) captures the factors that allow a given strategy to persist in a population. Thus it allows a clear way of predicting and explaining the existence of some traits. Maynard Smith and Price offer this definition: "Roughly, an ESS is a strategy such that, if most of the members of the population adopt it, there is no "mutant" strategy that would give higher reproductive fitness." (Smith and Price 1973, 15) Given this definition, we can expect drift and selection to gradually weed out almost all other strategies from a population aside from the ESS. Thus, if frequency dependent models apply to a given species-typical behavior, the evolutionary stability of the modeled behavior can offer a powerful explanation for its species-typicality.

The model I am interested in explains a variety of phenomena that are widespread in the animal kingdom. For instance, it explains why aggressive interactions regarding resource competition are rarely protracted and why the outcome of these interactions tend to be determined by prior ownership of a resource. Moreover, the model applies to a wide range of contests in the animal kingdom, ones that are used to determine ownership of various resources such as a territory. Specifically, the model applies to contests where nonlethal strategies are used. In these contests, the cost of an aggressive encounter builds up over time, and the disputed resource goes to the organism that persists the longest. Thus, game theorists call this kind of game the "war of attrition" (Maynard Smith 1974).

It turns out that there is no *pure* strategy for this game that is evolutionarily stable. A pure strategy in this case would be one in which an organism persists for a fixed amount of time, m, at each contest. No pure strategy will work because a competing strategy can be shown to have better expected payoffs regardless of the value of m (see

Smith and Price 1973). Rather the ESS will be a *mixed* strategy in which the length of persistence is drawn from a specific probability distribution, the mean of which is determined by the value of the resource and the cost of persisting. Specifically, the mean of the probability distribution is a duration of persistence the cost of which equals the value of the resource under dispute. In a population that consists entirely of this strategy, no pure strategy can invade. However, the expected value of this strategy is still only zero in a population in which everyone adopts it (see Maynard Smith 1974). The organism playing this strategy (in such a population) is unlikely to gain anything when the average cost of persisting and the average benefit of winning are summed up.

Maynard Smith points out that a better strategy would be to decide competitions with a coin toss. In a population dominated by the mixed ESS (described above), the probability that an organism would win any given contest is .5 anyway. So instead of wasting energy determining who by chance happens to persist longest in a given match, everyone would benefit if the contest was instead determined by coin toss. With such a scheme in place, no one would accrue the costs of persisting. By flipping a coin, we introduce an arbitrary *asymmetry* into the contest, and everyone is better off if the asymmetry is used to resolve contests by *convention*. The expected value of adopting a conventional strategy that determines contests by coin toss would be half the value of the disputed resource for each contest, which is far better than any strategy that ignores the coin toss (zero for the ESS that ignores the asymmetry). If we look to nature, there is an asymmetry that can be used in just this way: whoever found the disputed resource first, or in other words, whoever happens to "own" it. If all such contests are dyadic interactions, then on average, an organism will be the owner of the resource in about half of the contests in which it becomes involved. Thus, ownership can be used in the same way as a coin toss might be used. If a population of organisms were to decide contests in the favor of resource owners, this convention should have the same effect as deciding

contests by a coin toss. Game theorists call this the "bourgeois convention". The set of strategies that use the bourgeois convention to settle contests I will call "bourgeois strategies". Just like the coin toss strategy, an organism following the bourgeois convention can expect to get half the value of all the resources that it competes for in a population of organisms that follow the convention.

The success in evolutionary models of bourgeois strategies may help explain why owners of resources usually win fights in rodent species and in a variety of other species as well. It may also explain why marking behaviors, like urinating strategically at the boundaries of one's territory, are so common among mammals. Even in absence of strategic marking behaviors, animals will inevitably urinate and defecate on their territories at a higher frequency than they would elsewhere. Thus, a territory will often end up smelling like its owner, making smell a difficult-to-fake signal, or index, of ownership (Maynard Smith and Harper 2003). Given the reliability of this index, it is easy to determine which contestant in a territorial dispute is the owner of the territory. Thus, territory ownership is an unambiguous asymmetry that can be exploited to determine the outcome of contests.

Importantly, the stability of bourgeois strategies depends on ownership being backed up by force. The owner of a resource must play a "reserve" strategy of fighting for a specific length of time, in case the convention is not respected. Otherwise, a bourgeois convention will not be stable against a "mutant" strategy that ignores the asymmetry. The idea is that if the bourgeois convention is not backed up by a reserve strategy on the part of owners (e.g. if they were to relinquish the resource when an intruder attacks), then a certain range of mutant strategies can "call bluff" and win almost every contest with minimal cost in a population of bourgeois strategists. The set of bourgeois strategies that include a reserve strategy, I will call "bourgeois reserve strategies". Interestingly,

selection on the reserve strategy results in an ESS in which owners persist until the cost of fighting equals the value of the disputed resource (Rubenstein 1981).

Notice that in a population of bourgeois reserve strategists, the reserve strategy will never be observed (unless through some mistake in who is the owner). If everyone in the population respects ownership, then intruders will forfeit the resource to the owner before the owner plays the reserve. It follows that the motivation to play the reserve strategy cannot be prospective. If, as explained above, the stability of the bourgeois strategy depends on there being a fixed tendency to play the reserve strategy when the convention is violated, then the motivation to play the reserve strategy cannot depend on the prospective value, or anticipated reward for doing so. In a population of bourgeois strategist, there would be no anticipated reward for playing the reserve, because most organisms would never have played the reserve strategy and thus would have no reason to anticipate a reward for initiating the strategy. Thus, evolutionarily stable strategies for resource competition require that organisms play the reserve strategy in the face of immediate costs. Since the reserve strategy also imposes costs on the intruder in response to the intrusion, the strategy can lead to instances of temporarily spiteful punishment. It follows that it requires a retributive motive to implement it. That is, it requires a motivation to impose costs that is not motivated by the immediate material benefits of doing so.

In fact, the very structure of frequency-dependent selection lends itself to a similar conclusion. The bourgeois reserve strategy cannot develop by learning about the consequences of playing the reserve. This is because the factors that conduce toward the evolutionary stability of this strategy are not factors that are invariantly present in the experience of all organisms in a species. For instance, an organism with a bourgeois reserve strategy could be nested in populations consisting of combinations of innumerably many strategies. In some of these conditions, for instance in a population

dominated by the bourgeois reserve strategy, organisms will not experience any success at all with the reserve strategy. Again, this is because fights will not persist at all in a population of bourgeois strategists. However, evolutionary stability requires playing the reserve strategy against mutant strategies that attack owners and persist until the cost of fighting greatly exceeds the value of the resource. Mutants like this will do no better than bourgeois reserve strategists in a population consisting entirely of bourgeois reserve strategists, but only if reserve is played every time a mutant competes (or in other words, only if the population really consists of bourgeois reserve strategists rather than merely bourgeois strategists). In such a population, the mutant may win every fight for a resource, but it will take on considerable costs in half of its disputes, whereas the bourgeois reserve strategists will never take on costs for persisting (except in the rare encounter with the mutant) but will get the resource in about half of their fights. Thus, the bourgeois reserve strategy is only stable against this kind of mutant strategy because of the tendency to persist in fighting regardless of whether it has been successful in the past. In other words, learning from individual experience would not tend to converge on the bourgeois reserve strategy across all of the conditions required for its stability. Since learning would not reliably produce the relevant phenotype, it cannot be the mechanism by which the bourgeois reserve strategy develops.

This is a kind of "poverty of the stimulus" argument. If the motivation to play the reserve strategy depended on prospective processes, then it would require learning the contingency of reward on playing reserve.[11] In that case, there would be no reason to expect that individuals would end up using the evolutionarily stable strategy regarding when to play the reserve. Thus the bourgeois reserve strategy has to develop invariantly across variation in fighting success and it has to be non-prospective.

[11] This just follows from the definition of prospection.

Given the problems of interaction complexity and strategic complexity in the previous section, this is clearly a better explanation for why some organisms have a retributive motive. The fact that the model applies to organisms without prospection eliminates strategic complexity and limits the interaction complexity required to implement the strategy. The only capacities required of an organism are that it needs to identify situations of resource competition and to attack for a specific duration if the intruder does not retreat. While this is still a complex capacity, eliminating the interaction with prospective capacities leaves us far less mystified about how such a variant could have originated. Moreover, it is easy to see how this capacity could be built up from more basic capacities, for instance, the capacity to attack a conspecific with an unfamiliar smell. The model has the added benefit that it would explain why a retributive motive would be constrained by proportionality, since the optimal duration of fighting for the reserve strategy is determined by the value of the disputed resource.

However, this is not yet an explanation for why human beings possess a retributive motive. It could only offer such an explanation if this motive were preserved in *Homo sapiens*. So far, we have no reason to expect that it has been, nor do we have any reason to think that human (or even primate) resource competition is anything like the war of attrition. Likewise, it not yet obvious that human anger was shaped by these evolutionary forces. While I cannot offer such an argument here, I will make it plausible that the selection pressures captured by the war of attrition model are responsible for the structure of an anger-like motivational state of rats. In the following chapter, I argue that this motivational state is continuous with human anger.

Consider the instinctive patterns of territorial behavior in rats. These behaviors have been investigated in great detail using a resident-intruder experimental paradigm (For a review, see D. C. Blanchard and Blanchard 1984). In these experiments, resident rats (rats who have occupied a cage or colony for a few weeks) will attack unfamiliar male

rats introduced into their cage. The attacks of the resident and the defensive maneuvers of the intruder comprise sets of stereotyped behaviors. Each attack behavior of the resident is paired with a matching defensive maneuver of the intruder. The resident adopts a set of stereotyped postures and attacks aimed at biting the dorsal (back-side) surfaces of the intruder. On the other hand, the intruder adopts a distinctive set of stereotyped behaviors aimed at avoiding or blocking the resident's attempts to bite its back.

While these behaviors are certainly stereotyped, they are not brittle and reflexive. Rather, they have a high degree of flexibility. For instance, attacks of residents vary depending on the defensive strategy adopted by the intruder. Moreover, as discussed in the fourth chapter, the aggressive aim of biting the back can also produce instrumental behaviors (R. J. Blanchard et al. 1977).

These attacks can continue with little reprieve until the intruder rat has been either killed or removed, though the most likely consequence in the wild would be for the intruder to leave the resident's territory very early on in the interaction (Ewer 1971). Moreover, these attacks are different than the kind of attacks that would be directed against predators or prey (Panksepp 1971). Since the dorsal surfaces of rats are less vulnerable to injury and do not easily permit damage to vital organs, this form of attack does not seem to have a lethal function (D. C. Blanchard and Blanchard 1984; D. C. Blanchard and Blanchard 1988). Rather the attacks seem to function to drive off the unfamiliar male without inflicting lethal injury.

What scientists have discovered about these behaviors (the parity of the behaviors produced by resident and intruder rats, the stereotyped nature of these behaviors, their presence in socially naïve rats, and the coherent aim of the behaviors) indicates that they are produced by two different behavior programs: the confrontation and avoidance programs. Moreover, these programs seem to have evolved in response to

each other and the confrontation system seems to have been largely shaped by the demands of resource competition. Finally, like anger, the confrontation system is reactive. Resident rats will attempt to bite the back of an intruder rat independently of any anticipated reward or independently of any previously established contingency between back biting and reward (Eibl-Eibesfeldt 1961).

The war of attrition model leads to four predictions that are accurate concerning these patterns of conspecific aggression in rats. First, as I argued above, in order to implement the ESS, the tendency to attack an intruder needs to develop largely without the benefit of learning from social encounters. This is what scientist have observed. These patterns of confrontation and avoidance are present even in socially naïve rats, ones that have had no prior opportunity to observe or participate in aggressive or playful interactions with other rats (Eibl-Eibesfeldt 1961).

Second, given that territories have extremely high value and that the costs of persisting are low, we can predict that under the right conditions, residents will attack for long periods of time with little reprieve. This effect has been observed (Michael Potegal and TenBrink 1984; Michael Potegal 1992). These results also suggests that the motivation to attack may be fixed independently of any aggressive goal (other than the retreat of the intruder), as we would expect given the fixed duration of attack for the bourgeois reserve strategy.

Third, the ESS requires that organisms "respect" ownership, in the sense that it will attack if it is the owner and retreat if it is an intruder. This is exactly what is observed. A territory owner always attacks unfamiliar male rats under conditions in which a territory is valuable and the intruder's presence is threatening. Since the primary benefit of a territory for rats is to serve as a nesting site, territories become much more valuable when females occupy the territory. Accordingly, whenever females have occupied a territory, the male resident of that territory will attack unfamiliar male rats

after inspection and sufficient exposure (R. J. Blanchard et al. 1977). The only exceptions to this pattern are expected: when the owner's sense of smell has been blocked, when the unfamiliar male is prepubescent or has been castrated (K. J. Flannelly and Thor 1978). The former exception results from the fact that rats primarily identify one another by their sense of smell. It is obvious why we would expect owners not to attack young or castrated males. They do not pose a challenge to the territory owner given that the function of the territory is for reproduction.

Likewise, intruders always adopt avoidance behaviors and never win the confrontation. The only exception to this is when laboratory rats confront a wild rat intruder. The best explanation of this exception is that lab rats have been bred for hundreds of generations to be less aggressive, so the defensive attacks of wild rats are overwhelming to any lab rat. Nonetheless, there is good evidence that the same patterns of confrontation and avoidance obtain in wild rats and that the usual pattern is for intruders to flee the owner of a territory (Ewer 1971; D. C. Blanchard and Blanchard 1984).

Fourth, the scent of a territory plays an important role in determining the behavior of the resident and intruder. Residents make threat displays and then attack after sniffing the anogenital region of the intruder, and as mentioned before they do not adopt these postures when their sense of smell has been blocked (J R Alberts and Galef 1973).[12] Moreover, it is likely that the unfamiliar smell of a resident and its territory is a proximal cause of the intruder's adoption of avoidance behaviors. Specifically, it may decrease the intruder's confidence in aggressive encounters, whereas the pervasiveness of the resident's own scent markings may increase its confidence (D. Adams 1976).

[12] Likewise, the aggression of lactating female rats is diminished if their sense of smell is blocked (Kolunie and Stern 1995; Ferreira, Dahlöf, and Hansen 1987).

In sum, the ESS in the war of attrition model leads us to expect asymmetry of behavior based on ownership[13]. With modest assumptions it also leads to the prediction that this behavioral tendency will be elicited by olfactory cues and will lead to fixed behavioral tendencies of confrontation and avoidance in the resident and intruder rat, respectively. These are precisely the phenomena that scientists have observed. Thus, the war of attrition model offers a plausible explanation for the structure of the confrontation system, and it is plausible that the selection pressures captured by the model shaped this system.

3. Conclusion: Understanding Phylogenetically Ancient Adaptations

Importantly, the war of attrition model does not compete with the other explanations of punishment I have discussed. It is quite possible that the motivational states required for an ESS in the war of attrition were subsequently co-opted for different purposes at different points in evolutionary history. So a retributive motive for resource competition could have been co-opted to serve a reputational function or to support cooperation in large groups. Moreover, given its role in resource competition, it is easy to see how the motive could come to have effects in closely related domains of cooperation or deterrence, both of which concern resources in which an organism has an obvious stake. Organisms cooperate to acquire mutually beneficial resources and they are benefitted when other organisms are deterred from compromising those resources.

While it is compatible with these explanations of the function of the retributive motive, the war of attrition model is clearly a better explanation of how a retributive motive could have originated.[14] I argued that other selection models suffer from difficulties when used to explain the origin of retributive motives. The problem of interaction complexity makes the explanation of the motive less satisfying, since it is

[13] While there is some debate about whether ownership is always the asymmetry that determines behavior, other assymetries on which an ESS might be based are strongly correlated with ownership (see Maynard Smith and Harper for a detailed discussion).

[14] Though, recall that the retributive motive is not the primary explanandum of the other models.

difficult to see how a trait with such a complex relationship with other traits could have arisen due to simple changes in development or genetics. The problem of strategic complexity shows that these explanations are necessarily incomplete, since idealized models do not factor in all of the effects that the trait could have. The more primitive explanation of retributive motives offered by the war of attrition model helps to mitigate both of these problems; though I do not think it completely resolves them. This case shows that we may sometimes achieve greater insight into the origin of a capacity by looking further back in time at the more primitive selection problems from which it may have arisen.

There is another reason why it is beneficial to look at the patterns of aggression in less cognitively complex organisms. Due to the simplicity of their behavioral repertoire in relation to humans, it is a simpler matter to identify separate behavioral systems and assign them adaptive functions. While humans may have some of the same behavioral tendencies, they are masked by our enhanced abilities to regulate our emotions. Harmon-Jones and his colleagues put it this way:

> Basic emotions, such as anger, provide organisms with relatively complex and biologically prepared behavioral potentials that assist in coping with major challenges to their welfare (Panksepp 1998). However, these inherited behavioral tendencies exist only as potential ways of behaving in organisms with larger, more complex brains. Thus, although humans may possess the same emotional instincts as other animals, we may not be as controlled by the dictates of emotions and thus we have more choices (Panksepp 1994). That is, our emotions may be regulated and thus may not directly affect behavior. (Harmon-Jones, Peterson, and Harmon-Jones 2010, 61)

While one might be skeptical that any of the same underlying causes could be involved in human and non-human aggression (a skepticism I confront in the fourth chapter), it is

important to recognize that human psychology did not arise out of an evolutionary vacuum. It is not implausible that at least some dimensions of our emotional responses will be potentiated by phylogenetically ancient mechanisms. Nevertheless, we cannot just infer from superficial behavioral similarities that a specific psychological capacity of rats (e.g. the motivational states of resident rats) shares a common evolutionary history with a specific human emotion (Machery and Mallon 2010). The next chapter draws out this inference more carefully. In it, I show that the confrontation system in rats is connected with human anger through shared ancestry, making the war of attrition a plausible story about the ancient origins of anger.

Chapter 3
Angry Animals: Adjudicating Two Competing Homology Claims

Many philosophers of biology believe that the homology concept will be central to a scientific understanding of psychological kinds such as emotions. They argue that traits defined in terms of homology – features of organisms that derive from a trait of a common ancestor – have many of the desirable properties of natural kinds. For instance, they are homeostatic property clusters, which are projectable in the sense that they support extrapolative inferences (P. Griffiths 2006; P. E. Griffiths 1997; Assis and Brigandt 2009; Brigandt 2009). However, the recommendation that the mind be carved up into homologs by itself does not place useful constraints on scientific attempts to discern natural divisions within human and animal minds. This is because no one has adequately explained how homology claims can be subjected to hypothesis testing.

Consider an example. Many emotion researchers and theorists have suggested that anger is an innate adaptation that may be shared with nonhuman animals (e.g. Ekman 1999; Sell, Tooby, and Cosmides 2009). This raises the question of which behaviors might be manifestations of anger in non-human animals. Given the tight link between anger and aggression in humans, some aggression researchers propose that innate patterns of aggression in nonhuman animals are manifestations of anger. In other words, they propose that the system responsible for these phenomena is homologous with human anger, meaning that these complex traits are derived from a common ancestral trait.

As plausible as this may sound, there have been two incommensurate proposals along these lines, and there has been little progress in adjudicating between them. According to the ethological hypothesis, a repertoire of confrontational behaviors observed in "resident", territory-holding, rats reflects "an underlying emotional state" that is a primitive version of anger (D. C. Blanchard and Blanchard 1984, 17 see also; D.

C. Blanchard and Blanchard 1988; D. C. Blanchard and Blanchard 2003b). This behavioral repertoire is set in opposition to avoidance behaviors observed in intruder rats[1], which reflect fear. Moreover, the hypothesis holds that these two distinct emotional systems provide the best way of understanding angry aggression and fearful aggression in humans. Another proposal, the neurophysiological hypothesis is that human experiences of anger "emerge" from a pan-mammalian brain system that produces defensive behaviors that can be observed when areas within the ventral hypothalamus (among other areas) are electrically stimulated (Panksepp and Biven 2012; Panksepp 1998; Panksepp and Zellner 2004). These behaviors are set in opposition to predatory behaviors, which are neurally dissociable from the defensive behaviors. In other words, this hypothesis holds that there are two neural systems for aggression, and that one of them, the RAGE system, provides the primary neural substrate for human anger and is the proximate cause of "the feeling states and behavioral acts" (Panksepp, 1998, p. 14) distinctive of human anger. Moreover, the proponents of this hypothesis claim that we can best understand certain types of human aggression, impulsive and instrumental forms of aggression, in terms of the neural systems for defense and predation, respectively.

Importantly, these hypotheses are incompatible. Within the neurophysiological tradition, the neural dissociation between predatory and defensive aggression is the main reason to consider them fundamental, distinct categories of aggression. However, confrontation and avoidance behaviors do not exhibit this kind of clean neural dissociation (Siegel 2004, chap. 1). Moreover, the kinds of defensive aggression in rats produced by electrical brain stimulation is distinct from the aggression observed in

[1] The Blanchards actually refer to these behaviors as "offensive" and "defensive", respectively, and to the systems responsible for them as the "offense system" and the "defense system". For the sake of clarity, I use the terms "confrontation" and "avoidance" instead – as well as the corresponding "confrontation system" and "avoidance systems" – as terms of art to prevent the conflation of Panksepp's notion of defensive aggression with the Blanchards' notion of defensive aggression.

ethological research in the sense that it lacks features that are diagnostic of these forms of aggression (e.g. Kruk 1991). In other words, the aggression phenomena identified by these different research programs are behaviorally distinct and distinct neural mechanisms are responsible for them. As a result, they make incompatible inferences about what anger is and, more specifically, about which aggression phenomena are its manifestations. The bimodal classification schemes for aggression (defensive versus predatory and confrontational versus avoidant) that distinguish these respective phenomena are incommensurate.

While proponents of these hypotheses aim to identify homologies, there has been little progress in adjudicating between them. There are two reasons for this. One is the *target* problem: they have not carefully identified the human psychological trait that is the target of comparison. Another is the *evidence* problem: it is unclear how cross-species comparisons support homology claims. More specifically, it remains obscure how comparative evidence can play a role in adjudicating competing homology claims. While the issues pertaining to the target problem have received a good deal of attention in philosophy of biology, the evidence problem has been neither raised nor resolved. In this paper, I show a way forward by developing evidential criteria of homology and an evidential constraint on homology claims. I then apply these criteria and the constraint to the case of human anger and animal aggression to make it clear how hypothesis testing can proceed and which of the hypotheses above is supported by the evidence.

In the following section, I lay out some of the considerations that favor the use of *cladistic* categories (biological categories determined by common descent) in the introduction and modification of psychological and behavioral kind concepts (such as anger and aggression). I then argue that these considerations favor *basic human anger* as the target for the neurophysiological and ethological hypotheses. In section 2, I describe in detail the two hypotheses concerning human anger and the aggression

systems of non-human animals. While these hypotheses are homology claims, most of the evidence presented in their support is of questionable relevance to homology. In section 3, I say more about homology thinking. Homology thinking is a historical mode of thinking that explains similarities by appealing to common descent. To understand what kind of evidence supports homology, I point out a range of hypotheses with which it competes and set out the kind of evidence that favors homology over and above them. In section 4, I show how these criteria can be applied in the case of anger and aggression. A straightforward application of the criteria provides stronger support for the ethological hypothesis. Basic human anger has several similarities with the confrontational behaviors of resident rats, similarities that provide some evidence that these traits are a product of common descent. On the other hand, there is currently no evidence that the RAGE system uniquely corresponds with human anger. The similarities identified by the neurophysiological hypothesis hold not only with anger but also with other human emotions, such as fear. I conclude by highlighting the value of cross-species comparison for specifying psychological kinds.

1. The Identification Problem: Targeting Human Anger

First, I address the target problem by connecting it to a closely related problem, that of determining the meaning and reference of theoretical terms. This question looms large in a vast array of sciences, especially in areas of psychological research. For instance, it looms whenever we have questions like the following: what are emotions (or memory or attention or creativity or imagination, etc.)?; what different kinds of emotion (or memory, attention, creativity, imagination, etc.) are there?; and what is anger (or fear, semantic memory, episodic memory, visual attention, etc.)? At the outset, these questions are shaped by preconceptions about the domain of interest. Memory research, we might imagine, began with a vague question about human information storage. Nevertheless, it gradually became clear that there are quite different information storage

phenomena (e.g. short term and long term memory) that are unlikely to be explained by the same underlying process. The coarse categories of human memory or information storage then fall to the wayside as these refined categories take precedence in psychological research. This simple example illustrates that we do not just continue to lump phenomena together into the same category (e.g. generic memory or information storage) with which we posed a question, nor should we. Rather, our inquiry is and should be guided by the constraint that the phenomena of interest share common explanatory elements (e.g. short term memory processes), ones that support extrapolation from one instance of a category to another instance or to the whole category (e.g. short term memory processes across individual difference, across people in different cultures, across species, etc.).

However, the reference of theoretical terms, such as short term visual memory and anger or even charge and gravitation, is not only determined by the phenomena that their referents are supposed to explain. Reference is determined in part by the theory in which the term is embedded.[2] An examination of the history of theoretical terms like "phlogiston" and "caloric" – their use across different theories and their ultimate elimination – supports this claim (Kroon 1985). The reference of these terms (or lack thereof) is not merely determined by an explanandum phenomenon, but also by the distinctive explanatory project of the theory. Phlogiston was not defined simply as "that which explains combustion and related phenomena", otherwise it would have ultimately been identified with oxygen. Rather the reference of the word was overdetermined (see e.g. Nola 1980; Kroon 1985; Stanford and Kitcher 2000) because the use of the term was governed by the constraint that its referent explain combustion in *a particular way*, namely by exiting from burning materials (rather than by bonding with burning materials). In other words, the epistemic function of the term in testing theoretical

[2] See Griffiths (1997, vol. 1997, chap. 7) for a more detailed treatment of these claims.

claims demanded that the term's reference be overdetermined by addition of the theory's proposal as to *how* phlogiston explains combustion (Kroon 1985).

When these considerations are transposed to research on human anger, we find that research on it is so preparadigmatic that very little consensus has been reached even regarding the phenomena that anger will ultimately explain. It is even unclear which explanatory projects offer competing accounts of *how* anger explains a given set of phenomena. For instance, anger has been described in all of the following ways: as a syndrome that is caused by aversive stimuli and that includes an impulse to reactive aggression (Berkowitz 2012b); as an appraisal of an action as "a demeaning offense against me and mine" (Lazarus 1991); as a product of social construction that bears a complex relationship with aggression (Averill 1983); as a reaction aimed at making outcomes more equitable (Donnerstein and Hatfield 1983); as an evolved mechanism to regulate and recalibrate the dispositions of others toward oneself (A. Sell, Tooby, and Cosmides 2009); and as a mechanism for enforcing a specific kind of moral norm (Rozin, Lowery, and Haidt 1999).

This diversity of explanatory interests animates several worries about the reference of anger as a theoretical term. First, many function ascriptions are consistent with the folk usage of the word "anger" and its corresponding concept(s). Second, it seems unlikely that the explanatory interests underlying these diverse descriptions and function ascriptions will eventually converge. Third, it seems unlikely that each and every one of the theoretical entities that answer to the name "anger" in these accounts will end up referring to a projectable category (e.g. across species, cultures, and individual differences). Fourth, even if more than one of them do, they will probably not refer to the same projectable category. Thus, insofar as theorizing about anger is guided by the constraint that the explanandum phenomena be lumped into categories with common explanatory elements, it seems likely that we will eventually reinterpret or

discard many of the explanandum phenomena for these diverse accounts of anger. Even when we do, we may find more than one projectable category of anger.

This theoretical milieu poses two closely related problems. One is that it is unclear how research on anger should proceed. We cannot assume that these various approaches are referring to the same underlying entity or even that they mean the same thing by "anger", so it is unclear which theories stand in competition with one another. Another problem relates to the target problem. Looking to psychological research on anger by itself offers us little guidance in targeting anger for the purposes of cross species comparison. How then should we proceed? The argument of this section is that cross species comparisons, when properly understood, place important constraints on the kind of theory in which anger will enter as a term. I will briefly explain what these comparisons are supposed to accomplish and thus how these comparisons should be understood. I will then show how they help constrain the target of cross species comparisons with anger in human beings.

One thing that cross species comparisons do is to provide a way of testing claims about how common descent constrains psychological traits. Why should we be interested in such claims? One reason has to do with natural kinds, their function within scientific theories and the natural kind status of some biological categories, specifically ones that are defined in terms of common descent. Among other things, natural kinds refer to categories that are projectable, meaning that one can make reliable extrapolative inferences about unobserved instances of the category based on features of observed instances. The leading theory of natural kinds holds that the projectibility of natural kind categories (at least in sciences like biology) is explained by causal homeostatic mechanisms (R Boyd 1991). These mechanisms unite categories by providing a common explanation of property correlations between their instances. In other words, possession

of the homeostatic mechanism explains the reliability of extrapolative inferences between instances of a category.

Some biological categories, namely cladistic categories, are defined in terms of common descent, which is itself a causal homeostatic mechanism. That is, common descent explains property correlations because descendants of a common ancestor will inherit many of the same properties from the ancestral organism. Cladistics is an approach within biological taxonomy that classifies organisms into clades, or groups of organisms all the members of which descend from a common ancestor (or rather, a common ancestral population). Moreover, most of the organisms within a clade will usually possess a range of homologous traits, or trait that are shared because they were inherited from a common ancestor. For instance, all tetrapods possess forelimbs constituted by bones with a common structure, and they share this trait, in modified forms, because they all inherited this trait from a common ancestral organism. In other words, homologies are clusters of correlated properties that organisms share because of their common descent. Common descent is the homeostatic mechanism that explains our ability to extrapolate reliably from the properties of one tetrapod forelimb to those of another.

Thus, when cladistic categories like homology guide the postulation and modification of psychological kind concepts, we have some reason to expect that these kinds will converge on projectable categories. Within this framework, we can understand cross species comparisons as tests of homology claims. They can indicate whether the traits of different organisms derive from common descent. As such, cross species comparisons test whether these traits are projectable categories. If the same psychological kind is found in two species and if its presence in both is best explained by common descent, then there is a modicum of confirmation that the kind concept refers to a category with some degree of projectibility because of its grounding in common

descent. On the other hand, if the comparisons misfire even in closely related species, then there may be reason to re-conceptualize the relevant kind concept. This is because the grounds are undermined for one of the homeostatic mechanisms, common descent, that might support the projectibility of the category.

Consequently, if we are going to target anger for the purpose of cross species comparison of this kind, then the reference of anger needs to be fixed according to a theory that explains a set of phenomena in a particular way, namely by inheritance. Again, common descent explains shared features of organisms in large part because of inheritance from common ancestors. Likewise, if common descent explains part of the structure of anger, it follows that inheritance will factor into this explanation. Moreover, many of the phenomena of anger need to be explained by inheritance *as opposed to* culture or individual experience.

Fortunately, there is a significant family of theories according to which human anger phenomena are explained in terms of inheritance rather than in terms of culture or individual experience: basic emotion theory as articulated by the affect program tradition of emotion research. Following Darwin (1872), Ekman and others (Ekman et al. 1987; Ekman, Sorenson, and Friesen 1969; C. E. Izard 1971) pioneered this approach by finding pan-cultural, involuntary, facial expressions for some emotions, called *basic emotions* (e.g. anger, fear disgust, sadness, and joy). The universality of these facial expressions serves as an indicator that these facial expressions derive from inheritance (rather than cultural transmission or individual experience). These expressions then served as a scaffold for consolidating related phenomena that are also explainable by basic emotions. For instance, the coordination of different physiological response patterns for each involuntary facial expression of emotion (Ekman, Levenson, and Friesen 1983) provides some reason to expect that these physiological responses are governed by the same underlying entity as facial expressions of basic emotions.

Early on, Ekman (1977) suggested that basic emotions have genetic bases and dedicated neural substrates. However, it is often difficult to characterize the exact sense in which traits are genetically determined (see e.g. P. E. Griffiths and Machery 2008). This is because developmental processes that produce these traits often depend on environmental regularities (non-genetic determinants) for species-typical outcomes (P. E. Griffiths 2001; Ron Mallon and Weinberg 2006). However these details are to be worked out, there is a clear sense in which basic emotions are innate. We have basic emotions because of biological inheritance and not because of environmental regularities or cultural inheritance or learning from individual experience.

Culture and individual experience cannot explain many of the phenomena of basic emotions. For instance, involuntary facial expressions of basic emotions appear very early in development as does the capacity to recognize or respond to them (e.g. Carroll E. Izard, Hembree, and Huebner 1987). They appear spontaneously in those who could not have learned them from experience, as in those who are born both blind and deaf (Eibl-Eibesfeldt 1973). They are automatic in the sense that they come unbidden and are difficult to fake, suppress or control (e.g. Ekman and others 1971). Many of the facial expressions of basic emotions have homologous expressions in chimpanzees and other non-human primates (Chevalier-Skolnikoff 1973; L. Parr et al. 2007a). This means that the sets of muscle contractions involved in human facial expressions of anger naturally occur in the social interactions of these species. Finally, recognizable changes in vocal characteristics have also been found for each basic emotion, with anger being the most recognizable (Scherer 2003). Again, these features of basic emotion phenomena (e.g. early development, universal signals) provide evidence that the idiosyncrasies of individual experience play a limited role in shaping them. For instance, early development of emotional expression and recognition suggests that these emotional capacities develop faster and with more regularity than they would if they relied upon

general-purpose learning mechanisms in response to the information present in individual experiences.

In sum, the affect program tradition has amassed a good deal of evidence to indicate that involuntary facial expressions of anger (among other basic emotions) are part of a broader set of phenomena that are innate and adaptively tailored to deal with specific situations. While this capacity has yet to be fully specified or understood, there is some reason to expect that these phenomena will be explained in large part by common inheritance. Given that no other theory has offered an account of anger as an inherited trait with clearly articulated response components, the basic emotion of anger, or *basic human anger*, is the best target for any hypothesis that attempts to connect aggression phenomena in non-human animals to anger in humans.

2. Two Comparisons

Now that we have a clear target, we can consider each of the two comparative hypotheses proposed. The two hypotheses arise out of separate research traditions, but there are several similarities in the manner in which the hypotheses are derived (see Table 1 for a summary overview). Both traditions focus on a specific set of phenomena (e.g. the behavioral repertoire of a resident rat), with the goal of elaborating and explaining those phenomena in a particular way. Moreover, both traditions postulate a *system* that is responsible for the explanandum phenomenon. What does it means to postulate a system? Neither hypothetical system discussed below has been described at a sufficient level of detail to think of it as a mechanism. That is, there is not yet a list of the parts of the system or the interactions within it that would be capable of producing the phenomena in question (Machamer, Darden, and Craver 2000). Nevertheless, a complete explanation of the phenomena would probably involve a description of the mechanisms that constitute the system. Thus, "system" is a way of referring to whatever mechanisms explain the relevant phenomena in a certain way. For both hypotheses, the

postulated system is compared with anger. In order to elaborate these hypotheses, I need to say a little about each research tradition, its explanandum phenomena, and the system it postulates to explain those phenomena (see summary in Table 1). I will also present some of the main lines of evidence offered in support of each hypothesis, pointing out that this evidence has little to do with homology.

Table 1. Comparison of competing hypotheses

	Ethological hypothesis	**Neurophysiological hypothesis**
Research tradition	Ethology	Neurophysiology
Explanandum phenomenon	Resident male rats' confrontation of intruder rats	Attack elicited by hypothalamic stimulation in cats
Category of aggression	Confrontational attack as opposed to avoidance strategies	Defensive as opposed to predatory aggression
Postulated emotion system	Confrontation system	RAGE system
Adaptive function	Defend resources against challenging conspecifics	Defend against threats to self and resources
Sources of support	Distinct motivations and behaviors Similarities with anger: • sympathetic arousal (e.g. heart rate, blood pressure, piloerection) • hormonal changes • pre-programmed features (e.g. threat displays) • approach motivation • reactivity/ impulsivity • hostility? • Elicitors (e.g. challenge)	Reliable neurological dissociation Similarities with anger: • sympathetic arousal (e.g. heart rate, blood pressure, piloerection) • hormonal changes • pre-programmed features (e.g. threat displays) • reactivity/ impulsivity • hostility? • elicitors (e.g. threat)

2.1. The Ethological Hypotheses

The ethological hypothesis holds that human anger is comparable to an aggression system characterized in rats. This claim has been championed by Caroline and Robert Blanchard (2003a; 1988; 1984), David Adams (1976; 2006), and Michael Potegal (1994; M. Potegal and Stemmler 2010) and builds primarily on work in the *ethological tradition* of Konrad Lorenz (2003), Niko Tinbergen (1963), and Eibl-Eibesfeldt (1979). Nevertheless, there has been a great deal of neuroscientific research on the avoidance behaviors discussed below (e.g. Canteras 2002).

The explanandum phenomenon identified by this tradition is the confrontational, as opposed to avoidant, behaviors of rats in a resident-intruder experimental model.[3] In this model, a resident rat is put in a cage of which it is the primary occupant[4]. After a residency of some weeks, a male intruder (sometimes anesthetized) of a similar size is placed in the cage and left without a path of escape. The result is that the resident rat adopts a set of motivations and strategies that are oriented primarily toward confrontation, whereas the intruder adopts a set of motivations and strategies (including a distinctive form of attack) that are oriented toward avoidance.

These differences in motivation and strategy are diagnostic for classifying instances of aggression and for positing distinct systems. For example, the two behavioral repertoires include different kinds of threat and different kinds of attack. In resident rats, the attacks target bites at the dorsal surfaces of the neck and back, whereas the avoidance strategies of the intruder block the resident's access to the back and attacks focus on the resident's snout. The resident's attack is accompanied by a motivation to approach, whereas the intruder's attack, as well as other avoidance behaviors, are accompanied by the aim of escaping the confrontation or, failing that, avoiding (or deterring) bites to dorsal surfaces. In a more natural setting, these

[3] The same kind of attack can also be observed in female rodents defending their pups.
[4] An ovariectomized female is often added to avoid social isolation. The point is that the cage is the male rats territory, in that he has marked it with his scent.

motivations and strategies are identifiable but are usually adopted by both combatants at different times in a given encounter. So the upshot of the resident-intruder experimental paradigm is that it leads to a polarization of confrontation and avoidance strategies that it is not usually possible to observe in the wild. This model constitutes the Blanchards' contribution to this line of research, who following the work of others (Scott 1976), use a mixture of ethological and experimental methods to isolate and analyze these forms of aggression and to probe the motivations that underlie them. Another phenomenon that may involve confrontational aggression is the aggression of lactating female rodents toward unfamiliar adults (D. Albert et al. 1987).

Blanchard and Blanchard (1984, 1988) argue that the characteristics of confrontation and avoidance behavioral repertoires constitute evidence for the existence of two underlying emotion systems, which give rise to these behaviors: "[Avoidant] attack, and indeed the entire pattern of [avoidance] behavior, reflects fear. And [confrontational] attack, we maintain, similarly reflects an underlying emotional state, constituting at least a primitive analogue of what we call in humans 'anger'." (D. C. Blanchard and Blanchard 1984, 17) [5] That there are unified systems organizing confrontational and avoidant behaviors is evidenced by the directedness of each system toward accomplishing a characteristic set of aims:

> The more that [avoidance behavior] is examined, the more it becomes certain that there is some underlying element that provides order in the chaotic jumble of input and output factors. As one example of the complexities of this process, it can easily be demonstrated that a rat will react to a certain threatening stimulus in a specific situation on the basis of the features of the stimulus (discriminability, movement,

[5] While the Blanchards characterize their claim in terms of analogy and speak primarily of the adaptive function of these systems, they also advocate the use of these systems as animal models of aggression. But for animal models to be useful, the standard assumptions is that there needs to be a conserved mechanism shared by model and target organisms alike. In any case, the argument of this paper is that a homology claim is plausible for this aggression system.

> etc.). If a neutral stimulus is experimentally paired with this noxious stimulus, the animal usually begins very quickly to treat the neutral stimulus as a threat. However, rather than reacting to the now conditioned threat source as it did to the threat with which it was formerly paired, the animal will give [avoidance] responses that are appropriate to the relevant characteristics of the conditioned stimulus…Thus, what is learned is not a [avoidant] response per se, but the emotional reaction to the stimulus. This emotional reaction, in conjunction with stimulus and situational characteristics, then determines the precise response to be made. (D. C. Blanchard and Blanchard 1984, 18)

Not only do avoidance behaviors reveal an underlying element that determines the appropriate response depending on each situation, a specific set of aims unifies these behaviors. While each behavior in the avoidance repertoire (e.g. freezing, flight, attack, etc.) are distinct in nature, they are unified in that they each contribute to, and all are coordinated in pursuit of, the central aims of threat avoidance (from both conspecifics and predators) and risk assessment (see Blanchard and Blanchard 1988, 46-47). This kind of coordination around a set of aims would not be expected of a reflex system (see e.g. Konrad Lorenz 1957). While Blanchard and Blanchard make this case in less detail for the confrontation system, it is clear that the aim of confronting an intruder and biting its back coordinates the deployment of different strategies from the confrontational repertoire. In sum, the unity of confrontation and avoidance behaviors with respect to their respective aims is evidence for the existence of separate systems underlying these respective behaviors.

2.2. The Neurophysiological Hypothesis

Now consider the neurophysiological hypothesis. This is the claim that human anger is connected with a RAGE system that has been characterized primarily in cats. This hypothesis has been primarily championed by Jaak Panksepp (1998; Panksepp and

Zellner 2004; Panksepp and Biven 2012) and R. J. R. Blair (2012) and draws on a tradition of neurophysiological research that includes the work of Walter Hess, John Flynn and Alan Siegel.

The explanandum phenomenon for this tradition is called defensive rage and is displayed in the wild by cats that have been threatened. In these situations, aggression is accompanied by sympathetic arousal, piloerection, pupillary dilation, arched back, and hissing (Leyhausen 1979). Walter Hess (1954) was the first to elicit this behavior by stimulation of the hypothalamus, while Panksepp (Panksepp 1971), Siegel (2004) and others carried on extending this line of investigation.

Since this tradition draws from neurophysiology and neuroscience, this phenomenon is distinguished from another form of aggression by the fact that they have distinct subcortical neural substrates. Specifically, defensive rage produces a kind of attack that is distinct from predatory attack and can be elicited by stimulation from electrodes placed at a location in the ventral hypothalamus of cats and rats, a region also known as the Hypothalamic Aggression Area (HAA) (Siegel et al. 2010; Siegel et al. 1999; Siegel 2004; Kruk et al. 1983). This form of aggression is distinct from predatory attack (also called quiet-biting attack), which consists in stalking prey and pouncing. While these experiments are usually conducted in cats, some strains of laboratory rats display a similar form of predatory aggression against mice, and mice will exhibit this kind of predatory aggression toward crickets.[6] Not only is predatory attack unaccompanied by sympathetic arousal, pupillary dilation, or piloerection, the electrode placements that elicit predatory attack are also concentrated in different areas of the hypothalamus (specifically, the dorsolateral hypothalamus) from those that elicit defensive rage (Wasman and Flynn 1962). Sites that elicit defensive rage do not elicit predatory attack

[6] See e.g. Panksepp (1971). To my knowledge, no one has investigated whether predatory attack appears in any form in herbivores, so it is unclear whether this kind of attack applies to them.

and vice versa. The brain system that produces defensive rage behaviors is referred to by Panksepp (1998) as the RAGE system. According to Panksepp, the system includes all and only brain regions linked with defensive rage behavior, including specific locations in the amygdala, hypothalamus, and periaqueductal grey area.

To see what other aggression phenomena might be included in the category of defensive rage, we can compare it with another prominent classification system. Both Siegel and colleagues (2004; Weinshenker and Siegel 2002; Siegel and Victoroff 2009) and Panksepp (1998) mention Kenneth Moyer's (1976) classification of aggression in clarifying their conception of defensive rage. Whereas Moyer's classification posited six subtypes of aggression (predatory, inter-male, fear-induced, maternal, irritable, and sex-related aggression), Panksepp (1998) and colleagues (Panksepp and Biven, 2012) contrast their view with Moyer's, suggesting that there are only two forms of aggression (predatory and defensive aggression) that can be reliably differentiated on the basis of neurological data. Panksepp (1998, 2012) conceives of defensive rage as distinct from both predatory *and* inter-male aggression (which he claims does not have a primary neural substrate, see Panksepp and Biven 2012, chap. 4) while encompassing fear-induced aggression, maternal aggression, and irritable aggression.[7] Siegel and others also argue that it is useful to categorize aggression in this way (Siegel and Victoroff 2009).

It is important to note that defensive rage is not coextensive with the confrontational form of aggression isolated by the ethological tradition. The kind of aggression evoked by stimulating the HAA does not always have the same characteristics as a resident rats attack, nor does it have the same characteristics as the intruder's behaviors. Stimulation of the HAA in rats yields a mix of behaviors including not only

[7] I am not aware of behavioral evidence that connects defensive rage with these other categories of aggression. The connection is usually presented as conjecture.

those that are associated with confrontation, but also behaviors associated with avoidance, such as jumping attacks and threats (distress calls and teeth chattering, Kruk et al. 1984; Siegel et al. 1999). Some of the postures distinctive of the confrontational repertoire are notably absent from behaviors triggered by stimulation of the HAA. Moreover, stimulation of the same electrode placement can lead to confrontational or avoidant forms of attack depending on the level of stimulation (Kruk and van der Poel 1980)[8]. Siegel and others conclude that hypothalamic attack is a "...behavioural category in its own right" (Siegel et al. 1999, 364, see also Kruk & van der Poel, 1980). Finally, it is not clear whether and how the RAGE system can account for the motivational states that correspond with the confrontation and avoidance systems identified by the ethological tradition (see below for further discussion).

As a result of the forgoing evidence, there is some agreement that the behavioral repertoires identified by the resident-intruder model do not have independent neural substrates at the level of the hypothalamus (see e.g. Siegel 2004, ch. 1). We can conclude that these modes of individuating aggressive behavior are incommensurate. To summarize, this is because the RAGE system is responsible for a mixture of confrontation and avoidance behaviors but it does not appear to be responsible for some of the most distinctive behaviors of the confrontation system (e.g. lateral threat, approach motivation and lateral attack) and avoidance system (e.g. avoidance motivation) as understood by the ethological tradition. Thus, the systems responsible for these putative phenomena cannot be the same. While this incommensurability may indicate that one or both of these traditions have isolated spurious phenomena, I do not wish to defend any claim along these lines.[9]

[8] Consequently, it isn't out of the question that other brain regions could control confrontational and avoidant behaviors by modulating the rate of stimulation to the HAA.

[9] One can find hedged claims to this effect within both research traditions. See e.g. Adams (2006) for the claim that defensive rage is an artifact of stimulation techniques and Siegel (2004) for the claim that confrontational attack is not well-defined.

Importantly, some the evidence presented in support of these hypotheses has little to do with homology. For instance, the eliciting conditions for both aggression systems are thought of in largely adaptive terms. What is a threat but something that can reduce an organism's fitness directly or indirectly? What is a conspecific challenge but something that can compromise an organism's control over fitness relevant resources? Even the ethological hypothesis is sometimes presented explicitly in terms of analogy: "The view that natural defensive and aggressive behaviors of lower animals may provide an analogue to human emotions began with the pioneering work of Darwin (1872)." (D. C. Blanchard and Blanchard 1988, 44) Moreover, adaptational categories are sometimes used to identify classes of biological traits over and above homology thinking:

> Certainly these behaviors, and the circumstances in which they occur, appear to be highly adaptive and functional under normal conditions of life for individuals of virtually all mammalian species. Specifically, [confrontation] increases access to breeding females, with dominant males reproducing at a higher rate than subordinates... There is no evidence of any discontinuity in the adaptive values of either [confrontation or avoidance] for people as opposed to lower mammals. (D. C. Blanchard and Blanchard 1988, 48–49)

In these strands of aggression research, there is very little reflection concerning the kind of evidence that supports claims of behavioral and psychological homology as distinct from adaptation. Nonetheless, as our discussion of homology will show, adaptive function is a thin reed on which to base the identification of two traits. The very existence of homologies shows that the same trait can take on different functions within a lineage (e.g. tetrapod forelimbs are the same trait but with many functions), and that similar functions can be served by very different traits in separate lineages (e.g. bat wings and bird wings have similar functions but are not the same biological trait). In the following section, I will argue that, properly understood, the operational criteria of homology help

to clarify the evidence on which homology claims should be tested. The question that the operational criteria of homology help to answer is the following: are there similarities that provide evidence of common history between the two traits? If there are such similarities, this provides evidence of common ancestry that stands independently of evidence for shared adaptive function.

3. Homology and Its Competitors

Though the concept of homology is crucial to evolutionary thinking, it was conceived in the service of biological taxonomy prior to Darwin's time. Owen (1846) thought of homology as the sameness of an organ or structure in different organisms under every *form* and *function*. A common example of homology is the skeletal anatomy of the vertebrate forelimbs. The radius and ulna are bone structures that are common to bats, chimps, giraffes and manatees even though their forms and functions are dramatically different among these animals (see Figure 1). They can be more or less dense, thicker or thinner, longer or shorter, (though their spatial relationship to other bones of the forearms are preserved) and they can contribute to the different functions of swimming, flying, running and grasping in different organisms. So the radius and ulna are the same traits that occur in different animals, even though they have widely varying forms and functions within these various animals.

Now that evolutionary thinking has been integrated into biological systematics, one prominent idea about homology is that homology is a causal-historical concept(for a clarification and defense of this claim, see Ereshefsky 2012). Specifically, a homology refers to traits of various animals that derive from a trait of a common ancestor. In this way, shared ancestry is the common cause of each homologue, and this common cause explains similarities between the homologous traits. In the words of one biologist (with some help from Darwin), homology is "…grounded in 'descent, with modification,' a process that belongs to the past." (Rieppel 2005, 24)

Figure 1. The bones of some mammalian forelimbs. The radius (green) and ulna (red) are the same kind of bone, which takes on different forms and functions in different animals. Thanks to Valerie Wiegman for this illustration.

As a causal-historical concept, we can identify and refer to a homology without having or requiring detailed knowledge of the developmental and hereditary mechanisms that give rise to it, just as we can refer to a disease entity, such as measles or chicken pox, without knowing about its underlying causes (Putnam 1969). Nonetheless, we learn more about each homology as we learn more about its underlying causes, just as we learn more about chicken pox as we learn more about the virus that causes it.

Given the causal-historical nature of homology, there is a vast range of evidence that could bear on whether or not one trait is homologous to another. Some of the best evidence pertaining to homology comes from cladistics. If one has an independently established phylogenetic tree, one can look at the distribution of a candidate homology, or character, on that tree. If, for instance, the existence of a homology is more parsimonious than convergent evolution on one or more occasion, then there is a strong reason to think that a trait is homologous.

Nevertheless, before we can even look at the distribution of a character on a phylogenetic tree, we need to know how to identify the character in each taxon, which becomes a tricky matter when dealing with behavioral and psychological characters. For instance, knowing that humans have anger, that rats have a confrontation system, and that cats have a RAGE system does not determine which of these capacities are the *same* trait or character.

One way of addressing this problem is to use the operational criteria of homology. These criteria need not function as a definition of homology but instead we can use them to establish a consistent set of methods for ascertaining homologies and by extension, identical traits. The criteria of homology attempt to identify particular kinds of similarity, the kinds that are best explained by common history over and above a range of competing hypotheses. For any given similarity across clades, there are several hypotheses in competition with homology. One is that the similarity is only by chance. Another more probable possibility is that convergent evolution explains the correspondence. When a similarity is explained purely by convergent evolution, we have a clear case of analogy. Still another possibility in the behavioral domain is that similarity is explained by plastic developmental processes, particularly learning. In the clearest cases of plasticity, similarity can be explained entirely by convergent learning or development, perhaps shaped largely by task demands or shared developmental mechanisms.[10] The main competition is thus between hypotheses of homology, analogy, and developmental plasticity. Insofar as they function as evidence, the criteria of homology should help pick out similarites between traits that are explained by common ancestry and not convergent evolution or plastic developmental processes.

[10] See Brown (2013) for a detailed discussion of the difficulties (e.g. due to the plasticity and transformability of behavior) in applying the criteria of homology to behavior.

The most prominent criteria for homology were developed by Adolf Remane (1971) and can be deployed for this purpose. Consider first the criterion of *position*. The criterion applies to the radius and ulna because even with different forms and functions across different organisms, they retain their relative position to other bones of vertebrate forelimbs (humerus and the bones of the wrist). It would be highly unlikely for these characters to have evolved de novo in each of the different animals that possess it and yet to have the same relative position to other structures. Moreover, there is no shared function across the different animals which possess this character that would explain the correspondence. While corresponding position sounds like a spatial property, it is actually topological, and can include corresponding positions in temporal sequence or corresponding positions across cognitive architectures (e.g. "boxologies").

The criterion of *special quality* concerns "…shared features [that] cannot be explained by the role of a part in the life of the organism. The fact that in the vertebrate eye the blood supply to the retina lies between the retina and the source of light is a famous example of a 'special quality'." (P. E. Griffiths 2007, 648) The more complex a shared quality is, the less likely that they would have evolved independently. The location of blood supply to the vertebrate retina is both complex and non-essential (and even slightly counterproductive) given the functional role of the retina (what it is used for in the organism), so it identifies a correspondence that provides strong evidence that the various instances of this character derive from common descent.

Finally, the criterion of *intermediate forms* allows identification of homologous forms, A and C, because of the existence of one or more transitional states, $B_1 \ldots B_n$, between the two forms. In many cases, the homology between transitional forms, say between A and B_1 or between B_1 and B_2, is determined by applying the other two criteria. For instance, there are transition states between the bones of the mammalian inner ear and the bones of the reptilian jaw. We know this because the bones of the reptilian jaw

share the same *position* (relative to other bones of the jaw) as the bones of several intermediate forms, some of which share the same position as the bones of the mammalian inner ear.

While the examples so far deal straightforwardly with morphology or body structure, all three of Remane's criteria have also been applied to behavioral and psychological traits by ethologists (for overviews, see Ereshefsky 2007; Wenzel 1992). The following are some clear examples of how Remane's criteria have been applied to behavior. The position criterion, for instance, can be applied to temporal position in the same way that it is applied to spatial position. Accordingly, this criterion is satisfied when the behaviors of different organisms occupy the same position in a broader sequence of behavior. Wenzel (1992) uses the example of different tail movements in different species of *Tilapia* fish, which occupy the same position in the sequence of behaviors that comprise the courtship ritual.[11]

In the case of *special quality*, Alexander (1962) discusses several cases in which different species of crickets have identical songs and in which these songs are considered homologous. These songs have complex acoustical properties that constitute a special quality. This is because the acoustical properties are arbitrary relative to their function. That is, their function is to assist in locating and courting a mate, and this function could be filled by a number of different song patterns that would differentiate between local species and allow for collocation of mating pairs. Indeed, similarities and differences in song have been a primary tool for determining taxonomical relations between different kinds of cricket, grasshopper, katydid, locust and cicada. Importantly, when a similarity is arbitrary with respect to function, this provides evidence in favor of homology over hypotheses of both analogy and ontogenetic adaptation. This is because it is improbable

[11] One difficulty with deploying this criterion concerns the transformability of behavior from one species to another (e.g. Beer 1984). Whether behavior sequences in separate species are similar depends on how behaviors are segmented. This and other worries make behavioral homologies more tenuous as taxonomic distance increases.

that processes like learning or convergent evolution would lead to the same arbitrary quality when any number of other arbitrary qualities (e.g. acoustical patterns) could have satisfied the same function. Alexander (1962) and Ereshefsky (2007) both point out that different joint movements sometimes produce homologous cricket or grasshopper songs, demonstrating that different body structures can execute homologous behaviors. The patterning of songs can correspond in ways that indicate shared ancestry without the song being produced in the same way.

For my purposes, an important constraint on homology claims derives from the fact that some homologies are nested within other homologies. For instance, the class of tetrapod forelimbs is nested within the class of paired appendages. Thus, the forelimbs of reptiles, amphibians, mammals and avians are members of the homology class of tetrapod forelimbs, but they are also members of the more inclusive homology class of paired appendages, which also includes the pectoral fins of sharks and teleosts. While pectoral fins are homologous with instances of tetrapod forelimbs as paired appendages, the similarities between pectoral fins and tetrapod forelimbs do not provide evidence for homology at the less inclusive level of tetrapod forelimbs. Inclusion in this more specific class is indicated by bone structures that are absent in pectoral fins. These structures are due to modifications that occured subsequent to the divergence of tetrapods from teleosts, and that is why teleost pectoral fins are not included in this homology class.

As a result, some similarities only indicate inclusion in a broader homology class (e.g. paired appendages), whereas other similarities indicate inclusion in narrower homology classes (e.g. tetrapod forelimbs). In other words, some similarities (e.g. those between pectoral fins and forelimbs) only provide evidence for inclusion in broader homology classes (e.g. paired appendages rather than tetrapod forelimbs). It follows that, when evaluating similarities between traits, it is sometimes necessary to consider which homology class a similarity indicates.

From these considerations, we can derive an evidential constraint on homology claims. To see this, consider the correspondences between a human forelimb and a feline hind limb. The criterion of position is satisfied, because there are similarities between the parts (e.g. between humerus and femur). There are relations of homology between these traits. They are homologous as mammalian extremities and tetrapod extremities. Nevertheless, if we were to specify the homology class as one that includes human forelimbs but excludes human hind limbs, the similarity in question does not provide evidence for homology at this level. This is because there are no similarities between the human forelimb and cat hind limb that are not also shared between the human forelimb and hind limb. Thus, to provide evidence for relations of homology at the level of some homology class G (in this case, the homology class that includes forelimbs but excludes hindlimbs) as opposed to the more inclusive class, H (in this case, homology classes that includes forelimbs and hindlimbs), requires that some similarities between relata are not shared by traits in the more inclusive class, H. I call this an "evidential constraint" on homology claims.

While the examples so far deal straightforwardly with morphology or body structure, all three of Remane's criteria have also been applied to behavioral and psychological traits by ethologists (for overviews, see Ereshefsky 2007; Wenzel 1992). I suspect that what seems obvious concerning morphology might be easily confused concerning behavior or psychology. As a result, one could find evidence that psychological traits are homologous, but misidentify the homology class that this evidence supports. One way of doing so is to violate the evidential constraint above. I will argue that the neurophysiological hypothesis is an instance of this mistake. As yet, there is no evidence that the RAGE system identified by neurophysiological research is a member of the homology class that includes anger but excludes other human emotions.

This is because the hypothesis does not identify any similarities that are not shared with other human emotions. I spell out the details of this argument in the following section.

In summary, homology is a causal-historical concept, and homology thinking is a way of providing historical explanations for observed similarities between biological traits (or characters). The evidential criteria for homology can isolate evidence pertaining to this kind of historical explanation. In the following section, I show how the evidential criteria can discriminate between the two hypotheses laid out in section 2.

4. Which Kinds of Aggression are Manifestations of Anger?

Now we are in a position to evaluate the ethological and neurophysiological hypotheses. Recall that the two hypotheses focus on different sets of phenomena. The ethological hypothesis focuses on patterns of confrontational behavior of territory-holding, "resident" rats, whereas the neurophysiological hypothesis focuses on patterns of defensive behavior elicited by electrical brain stimulation. The ethological hypothesis lumps its phenomena together according to contrasting motives of behavior (confrontation versus defense), whereas the neurophysiological hypothesis lumps its phenomena together according to dissociable neural substrates of behavior (regions of the hypothalamus that elicit defense behavior versus distinct regions that elicit predation behavior). In this section, I show how homological thinking helps to adjudicate between them. Arguably, the available evidence supports the ethological hypothesis over the neurophysiological hypothesis. When applied to the ethological hypothesis, the criteria of homology identify similarities between human anger and the confrontation system that indicate a common ancestral origin (see Table 2 for a summary). The proponents of the neurophysiological hypothesis, on the other hand, have not yet identified similarities of this kind. Rather, the similarities observed between human anger and the RAGE system also hold in relation to a broader set of human emotions.

Table 2. Summary of criteria of homology applied to competing hypotheses.

	Ethological Hypothesis	**Neurophysiological Hypothesis**
Position	Sequence of behavior (autonomic arousal & signal -> attack)	Sequence of behavior (autonomic arousal & signal -> attack) – no evidence to favor a correspondence with anger as opposed to fear signals; *not a unique correspondence with anger*
Special quality	Back-biting attack shared between rats and macaques Morphological homology between facial expressions accompanying aggression in humans and primates	Facial expressions during HAA stimulation in humans?
Continuity of intermediates	Aggression paired with homologous signals across descendants (i.e. rats, macaques, chimpanzees and humans) of intermediates	HAA stimulation elicits urge to attack across descendants (i.e. cats, rats, and macaques) of intermediates – *unique correspondence with human anger not substantiated*

First, consider the ethological hypothesis. There are other aspects of the confrontation system that may share corresponding positions in sequences of behavior across the taxa that share a common ancestor with rats and humans. For instance, increases in blood pressure are correlated with confrontational aggression in rats and with anger and aggression in humans (Fokkema, Koolhaas, and van der Gugten 1995; Ekman, Levenson, and Friesen 1983; Levenson, Ekman, and Friesen 1990; Levenson 1992; J. E. Hokanson, Burgess, and Cohen 1963; J. Hokanson and Burgess 1962; J. E. Hokanson and Shetler 1961; Geen, Stonner, and Shope 1975; Tyson 1998; Gambaro and Rabin 1969). Moreover, piloerection, another indicator of autonomic arousal can accompany confrontational attack in rats as well as an approach-oriented aggressive behavior pattern in chimpanzees (R. J. Blanchard et al. 1977; Goodall 1986; L. a. Parr, Cohen, and Waal 2005). Finally, there is some indication that increases in plasma

noradrenalin accompanying provocation correlates with aggressiveness in both male rats and human males (Fokkema, Koolhaas, and van der Gugten 1995; Gerra, Zaimovic, and Avanzini 1997) In other words, there may be correspondence between several indicators of autonomic arousal that precede attack in non-human animals and that accompany anger in humans.

Perhaps the strongest pieces of evidence for homology, however, is a special quality that is shared by rats and stumptail macaques. Adams and Shoel (1981a) note that dominant macaques and resident rats both implement strategies aimed at accessing the back and biting it. In macaques, this behavior seems arbitrary with respect to the (probable) function of inflicting non-lethal damage on the subordinate. Macaques have a much larger repertoire of bodily movements than rats, many of which could serve the function of inflicting non-lethal harm (pushing, kicking, scratching, slapping, holding etc.). Thus, back-biting is a *special quality*, and the best explanation of this behavior may appeal to products of common ancestry. In other words, the reason that the attacks of both rats and macaques are aimed at biting the neck and back may be that they share a common ancestor with a corresponding aggressive strategy and perhaps similar motivational mechanisms for negotiating intraspecific conflict.[12] There is some evidence that human anger includes an impulse to approach and attack, but no one has demonstrated that the impulse is pan-cultural or species-typical.[13]

While Adams and Schoel did observe several facial expressions of subordinate macaques, they did not note any facial expressions that uniquely accompanied the attacks of a dominant macaque. However, in more ecologically valid studies of macaque behavior, macaques with higher dominance status do display facial expressions toward lower ranking macaques in aggressive encounters, expressions that resemble anger

[12] Adams and Schoel argue for homology by considering similarity in the dynamic of attack and submission across both species.
[13] See e.g. Carver and Harmon-Jones (2009); Baron (1971); Berkowitz et al (1981); and Pedersen et al (2011).

expressions in humans (Chevalier-Skolnikoff 1974).[14] Chevalier-Skolnikoff (1973) argues that two of these expressions are similar (utilizing homologous action units) across macaques, chimps, and humans. Some confirmation of these comparisons has been attained by comparison using a facial action coding system to quantify chimpanzee facial expressions (L. Parr et al. 2007b). Thus, there is *continuity across the intermediates* for some components of putative aggression systems across the common ancestors of these species.

Now consider the neurophysiological hypothesis. The problem is that the case for homology is incomplete. First, there is some evidence for correspondence that has *continuity across intermediates*: stimulation of the hypothalamus of cats, possums, rats and marmoset monkeys leads to similar forms of attack (Roberts, Steinberg, and Means 1967; Bergquist 1970; Panksepp 1971; Woodworth 1971; cited in Lipp and Hunsperger 1978).[15] However, ethical and practical considerations make it nearly impossible to obtain evidence concerning the effects of hypothalamus stimulation in humans. It remains uncertain whether it would lead to attack or to any of the other concomitants of human anger (e.g. experiences of anger, facial expressions of anger, or physiological changes associated with anger, as distinct from fear). Nor have any of these studies observed distinctive facial expressions that indicate continuity with human anger.[16]

---

[14] Chevalier-Skolnikoff calls these expressions "stare", "round-mouthed stare" and "open-mouthed stare".

[15] Delgado (1968) produced aggressive behaviors with electrical stimulation of the thalamus and cerebellum of chimpanzees and macaques. However, these brain structures are notably absent from the neurophysiological hypothesis and its descriptions of brain structures involved in aggression. Moreover, Delgado and colleagues did evaluate facial expressions. However, these facial expressions were not analyzed.

[16] It is compelling that in macaques, stimulation only results in attack under certain conditions (M. Alexander and Perachio 1973), some of which depend on whether the electrical stimulation occurs in the presence of a higher or lower ranking conspecific (attack being more likely in the latter case). Nevertheless, one cannot conclude from this that this form of aggression is of a piece with the aggressive syndrome which includes angry facial expression. It is quite possible that there are several forms of impulsive aggression that an animal might inflict only upon lower ranking conspecifics, including pain induced aggression, fear induced aggression or perhaps even disgust induced aggression. Neither is it obvious that any of these forms of aggression are of the same kind as angry aggression. By contrast, the work of Adams and Schoel (1982), and Chevalier-Skolnikoff (1973) describes a certain kind of confrontational or dominance-influenced aggression *with which angry facial expressions are associated*. The same is not true

There is some evidence that amygdala stimulation can produce feelings of anger (e.g. Hitchcock and Cairns 1973). This evidence is even bolstered by the fact that stimulation of the medial amygdala in cats can potentiate defensive behaviors elicited by electrical stimulation of the hypothalamus (e.g. Shaikh, Steinberg, and Siegel 1993). However, several other emotional experiences beside anger have also been reported as a result of amygdala stimulation in humans, including anxiety, guilt, embarrassment, jealousy, and a "desire for flight or escape" (which is more strongly associated with human fear, see Frijda, Kuipers, and ter Schure 1989). It seems that current evidence does not support a distinct localization of anger-like and fear-like feelings or behaviors within the HAA or in the other brain structures that make up the RAGE system (in cats or otherwise). Thus, the evidence from brain stimulation does not reveal a unique correspondence with human anger; one that is not also shared with other human emotions.

Second, consider the criterion of position. As with the confrontational attack observed in ethological work, physiological arousal and threat signals do occur prior to defensive attacks elicited by electrical brain stimulation. However, no evidence has been presented that either the signals or physiological arousal involved in these attacks are homologous with these components of human anger as opposed to human fear. Moreover, it seems unlikely that any such evidence will materialize.

This becomes apparent when we look closely at the work of Siegel and others on the HAA, which is cited as support for the neurophysiological hypothesis (Panksepp 1998, 2012). In fact, Siegel does not advocate the neurophysiological hypothesis, and in many cases makes claims that constitute evidence against it. In several places (including

---

of aggression elicited by electrical brain stimulation. The connection with angry facial expressions has not been made, nor has the behavioral syndrome been carefully circumscribed in ecologically valid conditions in most of the organisms in which it has been observed. Leyhausen (1979) has done this work concerning defensive aggression in cats, but he distinguishes this form of aggression from a confrontational form of aggression that includes a back-biting attack. I suspect that this latter form of aggression is more comparable to the confrontation system in rats (cf. Blanchard and Blanchard 1984).

Siegel 2004) Siegel compares defensive behaviors with a disorder known as Episodic Discontrol, which is marked by "…decreased impulse control – a characteristic common to defensive behavior – and altered perceptual states following stimuli evoking *anger, fear or rage*." (Siegel and Victoroff 2009, 213 emphasis mine) Indeed, many of the similarities that are noted between defensive behaviors and these forms of human aggression are characteristics of affectively driven behavior in general. Impulsivity is a characteristic of many kinds of emotion expression (see e.g. Frijda 1986), including fear, anger, sadness, and joy. Thus, the position criterion is not satisfied in a way that provides evidence for a homology between the RAGE system and anger that is not also shared between human anger and human fear.

By contrast, manifestations of the confrontation and avoidance systems in rats can be distinguished by quantifiable differences in the facial expressions of residents and intruders (Defensor et al. 2012), just as manifestations of anger and fear in humans can be distinguished by their distinctive facial expressions (e.g. Ekman and Friesen 1971). Moreover, resident and intruder rats have distinct forms of attack with distinct target sites. Thus, it is possible to distinguish *within rats* at least two different patterns of impulsive behaviors accompanied by distinct facial expressions. Moreover, some of the similarities between confrontation behaviors and angry behaviors in humans are not shared with fearful behaviors in humans or avoidance behaviors in rats. In other words, human anger and the confrontation system in rats do not violate the evidential constraint on homology claims (relativized to a homology class that only includes the emotion of anger) because they satisfy the evidential criteria of homology in ways that are not also satisfied by other emotions like fear. A related virtue of the ethological hypothesis is that it can distinguish angry aggression from the widely acknowledged category of *fear-induced* aggression (see esp. Moyer 1976). The same cannot be said for the neurophysiological hypothesis. I suspect that at least some of the phenomena

identified by the neurophysiological hypothesis reflect behavioral outcomes of fear, rather than (or perhaps in addition to) anger.

In sum, the case for homology between the RAGE system and anger (with respect to a category that includes anger but not other human emotions) may be similar to the case for homology between the cat hind limb and the human forelimb (with respect to a category that includes human forelimbs but not human hind limbs). The similarities so far observed do not evince a homology relation that excludes other emotions (especially fear), whereas the case for homology between the confrontation system and anger does evince such a relation.

5. Criticisms

I have argued so far that the confrontation system is homologous with human anger and that there is better evidence for this hypothesis than the threat-defense hypothesis. The argument depends on two claims that have been challenged in the literature on aggression in non-human animals. First, the confrontational tactics of rodents must actually be underpinned by a confrontation system. Second, the confrontation system of rodents must have phylogenetic continuity with anger in humans. The first claim has been challenged by Panksepp (1998) and others, (Panksepp and Zellner 2004; Panksepp and Biven 2012) who think that this form of attack is produced by the interaction of different emotion systems in the brain. They argue that intermale/competitive aggression is poorly understood and may arise from the interaction of SEEKING (a dopaminergic appetitive system) and RAGE systems rather than having an independent neural basis. Specifically, they suspect that confrontational behaviors are motivated by an urge or appetite to dominate and that this reflects the influence of the SEEKING system. Their assumption, which I will not question here, is that if there is not an independent neural substrate for confrontational attack, then the form of aggression is not a natural kind of aggression. Their conclusion seems to be that

the putative confrontation system does not exist, or that confrontational attack is not produced by a primary emotion system, or that confrontational attack is not a legitimate category, or natural kind, of aggression. Any of these claims would be problematic for the thesis defended here.

The argument depends on a questionable assumption and a dubious claim. First, it assumes that confrontational attack in rats is a phenomenon secluded to the category of intermale aggression. Certainly the paradigm instance of this form of aggression is territorial and intermale aggression. However, it tells against this interpretation that similar patterns of behavior can be observed in lactating mothers defending their young (D. Albert et al. 1987) and in other forms of aggression in female rodents (e.g. Syrian hamsters in Michael Potegal and TenBrink 1984).

Second, the argument depends on the claim that confrontational attack behaviors are motivated by an appetite for dominance. While it is likely that the *maintenance* of dominance relationships requires an (acquired?) appetite for aggression or dominance, it is not true that *all* dominance-related aggression involves such an appetite. For rats with successful histories of fighting, the opportunity to fight another male can be reinforcing (for a review, see M. Potegal 1979), and this may be a key factor in maintaining dominance. However, it is doubtful that the initial *formation* of dominance relationships requires any kind of desire for dominance.

Consider some of the details of intermale aggression in rats. It is only under certain conditions that male rats will fight each other for dominance. In the case of sibling pairs of wild rats (JR R Alberts, Jr, and Galef 1973) and also pairs of unrelated lab rats (C. Grant and Chance 1957), there is little evidence of formation of a dominance relationship between pairs. The development of such a relationship tends to require some kind of provoking conditions. For instance, the presence of females (Barnett and Stoddart 1969), increased numbers of rats (E. Grant and Chance 1958) or the social

isolation of one of the cohabitants (K. Flannelly and Lore 1975) is often required for fights to break out and for dominance relationships to be established between pairs. This evidence goes against the claim that dominance relationships arise as a consequence of an anticipatory or proactive "urge for dominance or competitiveness", which would predict the spontaneous formation of dominance relationships without provocation. Rather, it is more plausible that territoriality or competition first break out when there is some kind of perceived provocation (as is likely with socially isolated rats in which anxiety or social ineptitude might easily be interpreted as a provocation) or challenge (as when perceived "ownership" of a female is threatened by another male). Thus, initial establishment of dominance relationships may be caused by provocation, whereas in subsequent attacks from dominant organisms, Panksepp's SEEKING system may become more important.

What I am proposing is that these two systems may influence dominance relationships in different ways at different points across the development of these relationships. While I agree with Zellner and Panksepp that intermale aggression may involve a nuanced interplay between different systems that influence aggression, I think there is an important corollary to this insight: we need not suppose that all aggression related to "intermale aggression" has the same motivation. We need not suppose that *all* dominance-related fights between males involve a reaction to provocation nor that *all* such fights involve an appetite for dominance. If I am right, then this undermines Panksepp's reason for claiming that intermale aggression arises from the interaction of separate systems.

Remember that the case for homology between the confrontation system and anger also depends on phylogenetic continuity between these two systems. However, Albert and colleagues (1994) make an influential criticism of any attempt to identify confrontational attack with human aggression. They claim that confrontational

aggression is a form of hormone-dependent aggression and refer to a wealth of evidence that there is no direct dependency of human aggression on *serum* levels of hormones like testosterone (as there is in rodents). For instance, artificially increasing serum testosterone to supranormal levels does not increase various measures of aggression (see the metaanalysis in D. J. Albert, Walsh, and Jonik 1994, 409). Thus, they claim that the confrontation system is not an important cause of human aggression.

The primary response to this criticism is that a phylogenetic perspective does not predict a direct dependency of aggression on plasma testosterone levels. Rather, in some species the effects of testosterone on the *central nervous system* (CNS, e.g. on sexual and aggressive behaviors) and the effects of testosterone on the *reproduction and maturation* (e.g. in spermatogenesis and the development of secondary sex characters) are adaptive in different contexts. For instance, testosterone's plausible CNS function of increasing aggressive behavior is not always adaptive in ecological contexts in which its reproductive functions are (J. C. Wingfield, Lynn, and Soma 2001). In humans, sexual maturation (of secondary sex characters and testes) subsequent to rising testosterone levels occurs over a long period of time, and this rise in testosterone occurs well before mate competition would be appropriate (even relative to the era of evolutionary adaptedness). The opposite pattern can also be found in some species, in which aggressive behaviors are appropriate when physiological preparedness for reproduction is not. Wingfield et al (2001) point out (among a host of other examples) that in some species of birds (e.g. the sedentary song sparrow), territorial behaviors occur at life history stages that do not overlap with the breeding season.

There are many adaptive costs of testosterone that are consistent across several taxa. These include increased energy consumption, lowered fat stores, interference with pair-bonding, and decreased paternal care. Given these effects, it would be advantageous for the effects of testosterone on aggressive behaviors, reproduction and maturation to

be independently controlled in some species. Thus, Wingfield et al (2001) point out several putative mechanisms by which the influence of plasma testosterone on the nervous system could be mediated or by which testosterone levels in the CNS could be independently modulated (pp. 245-248). Independent control of these functions is predicted in some species because of the adaptive costs of maintaining high levels of circulating testosterone and the independent and context-dependent benefits of its various effects. Given the significance of these ecological considerations for the evolution of testosterone's effects, there is no reason to think that *direct* hormone-dependency (direct dependency on plasma testosterone as opposed to mediated dependency or independent modulation) is criterial for identifying confrontational aggression across lineages (among which different ecological conditions have prevailed in recent evolutionary history).

A more nuanced view about the dependency of human aggression on testosterone makes more plausible predictions. For instance, the Challenge hypothesis (J. Wingfield et al. 1990) predicts increases in testosterone *specifically in the context of status challenges that impact reproductive potential* , that pubertal increases in plasma testosterone (which function to bring reproductive organs to maturity) will not necessarily or immediately affect aggressive behaviors (because during maturation they are not yet adaptive for reproductive competition), and that paternal care will correlate with lower testosterone levels and decreased aggression. These predictions are largely confirmed in humans (see John Archer 2006 for a detailed review).

In sum, the most influential criticisms of the ethological hypothesis do not provide evidence against the identification of anger with the confrontation system in rodents. These criticisms involve mistaken or unsupported assumptions, on the one hand, assumptions about the motivation behind the formation of dominance relationships, and on the other hand, assumptions about the relationship between

plasma testosterone levels and confrontational aggression predicted by the ethological hypothesis.

6. Conclusion

The hypothesis of a behavioral homology between a trait of humans and rodents leads us to expect that corresponding clusters of behavior will be present in most of the taxa that share a common ancestor with humans and rodents (if not in most mammalian taxa). This includes not only members of rodentia and primates, but also those of lagomorpha (e.g. rabbits), scandentia (tree shrews), and dermoptera (flying lemurs). Across these taxa, there should be patterns of aggressive behavior that share a similar temporal sequence and special qualities. These clusters of behavior should include threat displays and piloerection prior to approach and attack. Especially in non-primates, when confrontational attacks are directed at conspecifics, they should be aimed at biting dorsal surfaces. There is already some evidence supporting some of these predictions in scandentia (Olsen 1969; Walletschek and Raab 1982) and, as already mentioned, in stumptail macaques and chimpanzees.

For some of the relevant taxa, there is increased visual acuity (in primates and scandentia) and a diminution of olfaction and olfactory receptors (in primates Preuss 2007). Thus, visual communication becomes more valuable, and facial expressions homologous to involuntary expressions of anger in humans should naturally occur in the context of confrontational attacks. This leads to the prediction that well-established phenomena of reactive aggression in primates, such as redirected aggression toward subordinates (Cheney and Seyfarth 1989; Aureli et al. 1992), should be accompanied by these facial expressions and also by piloerection and ANS activation.

Table 3. Summary of predictions and questions based on a phylogenetic approach to anger.

Predictions	Parsimonious distribution among taxa (i.e. primates, dermoptera, scandentia, rodentia, lagomorpha) Homologous facial expressions among primates associated with aggression (perhaps in redirected aggression)
Questions	What is the relationship between ANS activation and facial expressions across primate taxa? Do humans display piloerection when angry and aggressive? Is post-aggression relaxation response pan-cultural? Distributed across primate taxa? Is there a connection between anger and noradrenaline across taxa? Are there patterns of immunoreactivity across taxa?

While many of these predictions can only be evaluated by developing detailed ethograms of agonistic behavior patterns between both conspecifics and predators or prey[17], there is much experimental work that can be done as well on human and nonhuman animals. For instance, what is the relationship between ANS activation and angry facial expression in nonhuman primates? Do they bear a similar relationship to that found in humans (e.g. Levenson, Ekman, and Friesen 1990)? Do humans display evidence of vestigial piloerection response in a state of anger leading up to aggression (as reported anecdotally in Eibl-Eibesfeldt 1979)? What about related measures of skin conductance in humans (as observed in pilot data reported by Hubbard et al. 2010)?

[17] The evaluation of agonistic encounters with predator or prey is especially important when prey is of similar or greater size or when predators are of similar size or less. For instance, in the case of chimpanzees or gorillas (prey) and leopards (predator), confronting the predator is sometimes more beneficial than fleeing (see citations in Boesch 1991, 221). In any case, these are the conditions (greater benefit for confrontation than flight, due to relative size and speed) in which large ungulates are likely to display confrontational aggression against their predators (1974). In the case of chimpanzees who sometime hunt infant or juvenile baboons, adult male baboons will sometimes display ostensibly confrontational aggression against chimps who threaten their young (Goodall 1986, 286). In such a case, it may sometimes be adaptive for chimps (predator) to get angry at baboons (prey) and vice versa.

What terminates the physiological arousal initiated by anger or confrontational aggression? In humans, there is some evidence that consummation of angry and aggressive action sequences can reduce physiological arousal caused by anger (see Tyson 1998 for a review), is this relaxation response pan-cultural? Does a corresponding relaxation response occur subsequent to confrontational attacks in other species? Moreover, are there correlations between the production of hormones such as noradrenalin and confrontational aggression across the relevant taxa? Are there patterns of immunoreactivity subsequent to anger arousal across the relevant taxa? This is just to gesture briefly at the many predictions and questions generated by a hypothesis of homology, and to point out the many ways in which the hypothesis can be confirmed or put to the test.

While these predictions are important for the study of anger and aggression, the case study has broader implications for psychological categories. The lesson is this: homology thinking can provide independent criteria for evaluating substantive disagreements on – and for eliminating confusion about – the nature of psychological kinds. In absence of homology thinking, it is difficult to see how further knowledge about the RAGE system or the confrontation system would serve to determine which aggression systems in non-human animals are most like human anger. Indeed, this is probably one of the reasons why there has been little productive discussion between the advocates of the two hypotheses. Homology thinking in this case provides a set of independent theoretical constraints for identifying corresponding traits across taxa. In the service of this demonstration, I further developed some of the methods for thinking about homology (cf. Ereshefsky 2007; Ereshefsky 2012) as it applies to psychological kinds. This account helps to specify what kind of evidence supports homology claims, namely, identification of *unique* similarities *at the appropriate level* between traits;

similarities that provide evidence for common ancestry as opposed to common selective pressures.

Though counterintuitive from some perspectives, the concept of homology helps to clarify what counts as evidence for claims of character identity. Note that identical characters can have different character-states. For example, a human arm and whale fin are identical characters, because they are both instances of the tetrapod forelimb. Nevertheless, they are different character-states, because they represent different forms that this character can take. Homology thinking allows the identification of characters that take shape in dramatically different character states; it enables us to identify evolved characters that walk in the guise of dramatically different forms and functions. Anger is one such character.

Chapter 4

Emotional Action in Animals: Beyond Fixed Behaviors

Vengeance is a powerful and destructive motive. It is powerful because the behavior of otherwise reasonable people can be bent to this purpose; destructive because vengeance harms its target, sometimes without any benefit to the vengeful. Importantly, vengeful actions have a common purpose – one that we pick out with words and phrases like "retaliation", "retribution", "getting even", "balancing the scales", "settling scores" and "expressing hostility". Nevertheless, there are a variety of *means* by which this purpose can be achieved. There are no topological qualities shared by all vengeful actions (e.g. a specific mode of attack). Rather, their common purpose draws them together into a single category of action. Behaviors as widely varied as punching, stabbing, glaring, ignoring, stomping out of a room, and hiring a hit man can all count as vengeful because of their shared purpose of getting even and their directedness at the vulnerabilities of the agents they target.

Many explanations of vengeful behavior appeal to our animal nature rather than to our distinctively human qualities. The story usually goes something like this. Vengeance is an innate product of our evolutionary past (e.g. McCullough, Kurzban, and Tabak 2012) that defies straightforward rational explanation (e.g. John Elster 1990) and that instead finds its proximal cause in the emotional responses that we share with other animals (e.g. Barash and Lipton 2011). As I have argued, basic human anger is just such an emotion. Nevertheless, its plausibility as a proximate or developmental cause of vengeful behavior hinges on an important question: is it possible for *shared* emotions (shared with other animals) to cause the highly variable behaviors that constitute acts of vengeance and retribution? In animals, these emotions appear to produce only stereotyped, instinctive behaviors. If the same emotions also influence human behavior, why then do humans exhibit so many varying and incompatible responses when they

manifest shared emotions? The tension is between the variability of emotional behavior in humans and the specificity of emotional behavior in animals. In this chapter, I focus on a special case of the tension. I take a step toward resolving the apparent tension between the claim that anger is a shared emotion and the claim that it can motivate behaviors aimed at retribution or retaliation. To do this, I argue against a widely and casually held view concerning shared emotions, that in animals, they do not influence purposive behavior, or *action*.[1]

In the first section, I lay out a set of individually plausible but jointly implausible claims about basic emotions together with some of their presuppositions. Some shared emotions, like anger, are continuous with the emotions of nonhuman animals in humans, they motivate purposive behavior; yet in nonhuman animals, these emotional states seem only to cause stereotyped behavior. Much of the work of this section (and of this chapter as well) is to clarify when behaviors are purposive and when emotions motivate behavior. In the second section, I look at some empirical evidence, which demonstrates that angry behaviors in rats clearly are purposive, despite their apparent stereotypy. In the third section, I draw out the implications of this claim for a widely but casually held view in emotion theory. I conclude by considering some of the reasons that we might expect the behavioral effects of anger to be even more flexible in primates and humans,

[1] I use "purposive behavior" and "action" interchangeably. Some philosophers reserve the word action for planned behavior or behavior mediated by conscious knowledge of what one is doing. I do not use it in this way. There are clearly interesting behavioral phenomena that fall between this highly intellectualized notion of action and mere passive movements.

While "purposive behavior" is more cumbersome, it is sometimes helpful to contrast it with non-purposive behavior, emphasizing that they are both *behaviors*, distinct from passive movements (as when someone moves because they are pushed) and perhaps also monosynaptic reflex movements (though see Burge 2010; Dretske 1991, chap. 1; Millikan 1993 for a range of different notions of behavior). It is also helpful to have a way of talking about behavior in isolation from its purposiveness, as when one asks whether a specific behavior is purposive or not. I also use "purposive behavior" in preference to the more familiar "goal-directed behavior". The notion of a goal can be misleading, because some would understand a goal as an *explicit* representation of a behavior's end state *as a desired outcome*. I do not want to seem to beg any questions against those who think there can be implicit representations of the aim of a behavior (e.g. Frijda 2010).

and I consider recent evidence suggesting that anger in humans includes a biological disposition toward reactive aggression or retaliation.

1. Shared Emotions and Purposive Behaviors: A Soft Paradox

I begin by articulating what I call a "soft paradox" or a set of individually plausible but jointly implausible claims:

1) Some shared emotions are continuous between human and nonhuman animals with regard to their motivational structure.
2) In humans, these shared emotions motivate purposive behavior.
3) In animals, these shared emotions only cause stereotyped behaviors, which are not purposive.

This soft paradox has an important set of presuppositions. It only makes sense to talk about the continuity or discontinuity of emotions across human and non-human animals if these presuppositions are true: there are basic emotions; some basic emotions are innate adaptations (as I argued in the introductory chapter); and some basic emotions are *shared* with non-human animals (as I argued concerning anger in the previous chapter). It is from these presuppositions that the paradox arises.

To get in the grip of the paradox, suppose that for some shared emotions, such as anger and fear, there is no discontinuity in their motivational structure across human and nonhuman animals; that differences between the human and nonhuman forms of these emotions are in degree and not in kind. In animals, emotions like these are thought to cause only stereotyped behaviors, such as threat displays or tonic immobility (a view I discuss further in section 3). But if this is true, then it is difficult to see how shared emotions could directly cause flexible, purposive behaviors in humans (e.g. skilled evasion, revenge, and retaliation). Since stereotyped behaviors in animals do not seem to be flexible or purposive, it is difficult to see how the mechanisms responsible for these inflexible behaviors could be gradually modified (without the addition of novel

components) over the course of evolution to motivate the kind of flexible, purposive behaviors that constitutes human actions of revenge and retaliation.

In this paper, I argue against the claim that shared emotions in animals only cause non-purposive behaviors. Several distinctions are key to understanding this paradox: the distinction between evolutionary continuity and discontinuity, the distinction between purposive and non-purposive behaviors, and the distinction between actions motivated by emotion and those merely influenced by emotion (perhaps indirectly). In the remainder of this section, I clarify these distinctions.

1.1 Continuity and Discontinuity

First, consider continuity. A claim of continuity concerning anger would go beyond the mere claim that human and nonhuman anger are homologous, or derived from the same ancestral trait (a claim I supported in chapter 3). For instance, we might suppose that both human and nonhuman anger derive from a common ancestral trait, but that the human form of this trait underwent a fundamental change, perhaps involving the addition of a novel component. To claim that human anger is continuous with its nonhuman forms is to deny such a possibility. It is to claim that anger across human and non-human lineages derives from an ancestral trait without the addition of novel components.

Discussions of continuity versus discontinuity between different lineages arise most prominently in discussions of evolutionary novelty. In these discussions, novelty is sometimes thought to require non-homology (e.g. Brown 2013). This is only partly true. As I pointed out in the previous chapter, homology classes form nested hierarchies. Morphological units as wide ranging as human forearms and shark pectoral fins are homologous qua paired appendages, but within the category of paired appendages, there are mutually exclusive homology classes which constitute different states that the character *paired appendage* can take. For instance, paired appendages can be

cartilaginous or bony, and the class of bony paired appendages includes coelacanths pectoral fins and human forearms, but not shark pectoral fins. The evolution of bony paired appendages was an evolutionary novelty with respect to the broader class of paired appendages. From the moment it appeared, a new, nested homology class arose within the broader class of paired appendages. It is not as if, at that moment, it ceased to be homologous with other paired appendages. Rather, the newer state of this character is only non-homologous with the shark pectoral fin qua *bony* paired appendage. In other words, the evolutionary novelty of bony paired appendages established a new homology class within the broader one, a class that excludes some members of the broader class.

In light of this and the previous chapter, the suggestion that hominid anger is discontinuous with anger in rats amounts to the claim that anger in one of these lineages took on novel components, thus establishing a newer homology class that excludes the other. This would be consistent with the claim that the characters are homologous qua anger but non-homologous qua hominin anger or qua rodent anger.

For my purposes, the plausibility of discontinuity concerning the behavioral influence of emotion will depend on whether emotions have qualitatively different influences on action in human and nonhuman animals. There is little reason to posit discontinuity of influence over action if it turns out that human and nonhuman anger alike do not directly cause purposive behavior, or if human and nonhuman anger alike cause purposive behavior.

1.2 Purposive behavior

Now consider the distinction between purposive and nonpurposive behavior. In this subsection, I say more about what it means for a behavior to be purposive, and I briefly argue for two conditions under which we are justified in inferring that a behavior is purposive. Later, this will provide a principled basis on which to claim that emotional behavior in rodents is purposive.

As a first pass, purposive behavior is behavior caused (in the right way) by the informational and motivational states of an agent.[2] [3] When we say that a behavior is undertaken for a purpose, we are attributing a purpose *to the agent or organism* as well as some information about how to fulfill it, both of which play a causal role in the production of the behavior. Why "informational" and "motivational" states as opposed to traditional notions of belief, desire, and intention? I think these notions are not particularly clear or helpful for my purposes, and I want to avoid getting lost in disputes about how to understand these terms.[4] Moreover, I think there may be motivational states other than desires (e.g. urges, fears, embarrassments, appetites) and informational states other than beliefs (e.g. perceptual and motor representations, representations of affordances, etc.) that can cause purposive behavior. The view I take in what follows is that the purposiveness of behavior does not depend on what *type* of informational and motivational states cause behavior but on whether behavior is caused by these *representational* states *at all*. The question I will focus on below concerns when we have

[2] However, see Sehon (Sehon 2007; Sehon 1994)for an argument against understanding action in terms of causation. Sehon and others (Frankfurt and Frankfurt 2014) argue for a class of teleological explanations, claiming in addition that these explanations cannot be given a causal analysis. The plausibility of such claims seems to depend on a particular conception of causation that no longer holds sway. By contrast, if one holds a manipulationist account of causation, teleological explanations have a straightforward causal construal. Whether or not an organisms engages in a certain class of actions causally depends on whether or not it has a specific kind of goal. On a manipulationist account of causation, this just means that if we were to intervene to change the organism's goal, then its behavior would change in certain ways. Of course the teleological action theorist could argue that the concept of a goal cannot be analyzed in terms of causation. Nevertheless, the causal theorist could similarly deny that beliefs and desires can be given a causal analysis. In fact, a claim like this may fall out of Davidson's holism. If so, then whether or not a theory is a causal theory of action does not obviously depend on whether the variables in its causal explanations can themselves be given a causal analysis.

[3] The parenthetical "in the right way" is often inserted as a nod to the intractable problem of deviant causation (Davidson 1963). A version of this problem also plagues teleological theories of action (Mele 2000). I do not have space here to deal with this problem, nor is it necessary to do so for my purposes.

[4] Specifically, there is debate about what beliefs, desires and intentions are, sometimes accompanied by arguments that nonlinguistic animals cannot have beliefs and desires (Davidson 1986; Gauker 2003; Schroeder 2004; Railton 2012; Pacherie 2006). Interestingly, some researchers in animal learning do use the concepts of belief and desire to distinguish animal behaviors that are goal directed from those that are not (Balleine and Dickinson 1998; though for criticisms, see Carruthers 2004; Sterelny 2001). It is an interesting question how one would reconcile this usage to standard philosophical views about beliefs and desires. Nevertheless, it is a question that I cannot resolve here, nor is it necessary to resolve it for my purposes.

good evidence that behavior is caused by informational and motivational states, or equivalently, when we have good evidence that a behavior has a *psychological explanation*.[5]

Of course, we can interpolate the notions of information, motivation and purpose into almost any system. Famously, we can interpret the behavior of a rock as being guided by the purpose of being at the bottom of a hill (or for Aristotle, at the center of the universe); we can interpret thermostats as possessing information about the current temperature and a motivation to keep the temperature at 75 degrees (or whatever temperature it is set to); and we can predict the autumnal defoliation of deciduous trees by attributing beliefs about the impending winter. If representational states can be applied to almost anything, this explanatory practice will not be particularly helpful in distinguishing between purposive and non-purposive animal behaviors. So we need some way to regiment the practice of psychological explanation if it is going to be of any use.[6]

1.2.1 Regimenting the practice of psychological explanation

I think the mishap in the "explanations" above is that they conflate three very distinct explanatory projects, each with its own explanandum – or phenomenon to be

[5] Notice that I did not say "when behavior *seems* to be caused by informational and motivational states, or when we *seem* to have good evidence that a behavior has a psychological explanation". To me, the absence of "seems" marks the difference between folk psychology and psychology proper. This relates to another reason that I want to avoid disputes about the nature of belief and desire. These philosophical disputes often turn on folk psychological considerations, for instance how we *talk and think* about beliefs and desires. I am more interested in empirical considerations. For instance, what are the empirically observable phenomena that beliefs and desires are necessary to explain, and what do these states have to be like in order to produce these phenomena?

[6] The following discussion will be reminiscent of Dennett's work on the "intentional stance". This is no accident, because the issues I am concerned with are closely related to those of concern to Dennett. The main difference between my approach and Dennetts is that for Dennett, the unique predictive power and predictive success of the intentional stance is what justifies positing beliefs and desires (cf. Viger 2000). On my view, the reason to posit informational and motivational states is that there is a further phenomenon to be explained beyond what Dennett would call the physical stance and the design stance. I am interested in a more straightforward realistic construal of psychological explanation, and one that does not directly hinge on an inference from predictive power and success to the existence of theoretical entities. See Van Fraasen (1980) for an influential critique of this inference. It is difficult to tell whether Dennett makes this inference, but this is the most natural way of interpreting him if he is a realist of any kind.

explained. The explanatory project that I call psychological explanation only makes sense when directed at one of these explananda. To identify the right one, it helps to appreciate two different errors – the *animistic* error and the *Quixotic* error – that one makes when psychological explanations are applied too liberally. An appreciation of these errors marks the difference between the three different explananda.

One error is the *animistic* error of attributing purpose to the movement of a rock down a hill. Certainly, we can make fairly accurate (though not particularly precise) predictions about the movement of a rock by attributing a "desire" to move as close to the center of the earth as possible and a "belief" that a certain path down a hill is the path of least resistance. The mistake here is to think that once we have a physical prediction or explanation of the rock's movement (in terms of its mass, position, velocity, and the forces acting on it) there is some further phenomenon to explain. Only if there were some further phenomenon would there be any point in positing other theoretical terms to explain its trajectory, regardless of their functional role.

In the usual case, there is no such further phenomenon, but it would be a mistake to say that there never is anything further to explain, even where rocks are concerned. Consider the case of Indiana Jones. His foot hits a tripwire, and suddenly there is a massive boulder hurtling toward him. Even if one has a complete physical explanation of the rock's movement (e.g. explaining how the force exerted on the tripwire transferred to whatever mechanism was holding the rock in place…), there is still something else to explain. Why was the rock set up to move exactly when Jones walked into its path? To ask such a question is to inquire about the *structuring cause* of an event or process (Dretske 1991, 42). This further question inquires into the events or process that caused *the tripwire to* cause the rock to hurtle toward Jones.

Of course, there are different kinds of structuring causes. Sometimes a process is structured by design. Thermostats are set up to ignite the furnace when the temperature

dips below a set point, which has the effect of raising the room temperature. The reason why they do so is explained by the intentions of their designers. Other times, processes are structured by natural selection (or some other selection process). When an egg (or even a white cube of the right size) falls out of the nest of a graylag goose, the goose will engage in a species-typical, stereotyped movement that usually has the effect of bringing the egg back into the nest. Nonetheless, a graylag goose will slavishly carry on this movement even when the egg is snatched away (by an experimenter) well before the behavioral sequence is completed. Natural selection is almost certainly responsible for this movement. Graylag geese perform this movement today because their ancestors who performed the movement were more successful at raising their young to reproductive maturity than geese that did otherwise.

In all these cases, one should avoid the *Quixotic* error of crediting either the rock, the thermostat or the goose with structuring the relevant process. By contrast, behavioral processes are sometime structured by an organism itself rather than by some external process of design or selection. When I walk over to the refrigerator to get a beer, it is I (and not, say, the evolutionary forces that shaped me) who structures the behavioral process. In this case, psychological explanation becomes appropriate because the process is structured by motivations and means that belong to me in some sense.

While beer retrieval and egg retrieval are clear cases of agential structuring of behavior on the one hand and selectional structuring on the other, a critical difficulty remains. How can one tell the difference in more borderline cases? When are we justified in crediting the agent with structuring its behavior to achieve its own purposes and with means of its own devising? Here, I can do little more than describe some of the conditions that evince purposive behavior and gesture at why these conditions support the claim that an agent itself structures its behavior.

1.2.2 Behavioral plasticity and the explanatory role of motivational states

I believe that two conditions are jointly sufficient to infer that behavior is purposive, or that it has a psychological explanation. One mark of purposive behavior is its flexibility. Such behavior can be adjusted in various ways to bring about a certain outcome or end state, which is the hypothesized purpose of the behavior. Philosophers and behaviorists both have tried to capture this with the notion of *plasticity*.

Behavior is plastic when the organism can reach an outcome or end state in a number of different ways, some of which may be novel (to the organism) ways of causing that outcome.[7] If we imagine a rat swimming across a stream to get a piece of cheese, the behavior is plastic if we have evidence for a set of subjunctive conditionals concerning the rat's behavior: "'If the food were further to the left, the rat would be swimming further to the left', 'If the food were not on the far side of the river, the rat would not be swimming across it', and a host of others, more or less specific, affirmative or negative" (Woodfield 1976). These plasticity-conditionals are a way of describing the sort of flexible behavior adjustment in relation to an end state that is characteristic of purposive behaviors.[8] The counterfactual nature of these conditionals highlights the fact that psychological explanation does not only aim to explain what the animal *actually* does but also explains what we have good reason to believe that the animal *would have done*.[9]

[7] One obvious difficulty is individuating different ways of bringing about an end state in a principled way. This is a methodological difficulty that is best left to the ethologist and animal behaviorist. While there may be borderline cases when it is difficult to tell whether a means to an end is the same as one the organism has pursued before, there are also clear cases, like the one I discuss in the following section. Moreover, novel means for a given end are particularly clear cases.

[8] Plasticity can also be manifested in the persistence of behaviors. A behavior is persistent when various behaviors (appropriate to a given end state) continue until the end state has been reached. While it is sometimes treated separately, I think persistence is either special case of plasticity or it is an attempt to capture exchangeability (discussed in the following heading). If behavior does not persist in a range of conditions, then this will either limit the plasticity conditionals for which we have evidence or it will show that a given kind of behavior (that could be used to achieve a given end state) is not exchangeable across a range different end states.

[9] Importantly, a physical explanation of what an animal actually does could leave us wondering what an animal would have done on some of these counterfactuals. This is part of what it means to say that there would be a further phenomenon to be explained even if we had a complete physical explanation of actual behavior.

If an organism is only able to bring about an end state through one element of its behavioral repertoire, then we have little basis on which to credit the organism with guidance toward the end state (especially if all other members of the species are also capable of using that element to bring about that end state). If the graylag goose can only retrieve its eggs using a single type of movement, then we have less reason to think that the goose itself is trying to bring about the relevant end state and still less reason if the movement continues when it is no longer contributing to that end state. In that case, it seems unlikely that the goose itself has selected that behavior as a means to its own end.

By contrast, being able to bring about an end state in more than one way (especially if novel) provides some evidence for purposiveness. Why does plasticity give us reason to credit the organism with structuring its own behavior? Plasticity uniquely supports hypotheses that postulate an internal motivational state guiding behavior, as opposed to a range of (reasonable) alternative hypotheses. A paradigmatic behaviorist hypothesis, for instance, would try to explain behavior in terms of reinforcement relations between sensory and physiological stimuli and behavioral responses. Another alternative hypothesis would be that behavior is triggered, perhaps innately, by a highly constrained set of stimulus conditions. Call this an "ethological hypothesis". If behavior is sufficiently plastic, then hypotheses of these sorts will have difficulty predicting all the plasticity conditionals without adding in ad hoc assumptions.[10]

The reason is that both kinds of hypotheses explain behavior in terms of lawlike connections between inputs and outputs. Imagine the massive disjunction of lawlike generalizations that would be necessary to predict the input-output relations that

[10] Of course, the goose's behavior might *seem* to be plastic with respect to egg-retrieval under the following conditions. We might imagine being able to train a goose to use its feet to push an egg into a nest. However, it is very unlikely that this kind of behavior would be plastic with respect egg retrieval. Instead, it would likely be a plastic response for increasing the frequency of positive reinforcers (or diminishing the frequency of negative reinforcers). The point is that plasticity is relative to a specific end state, and that the end state toward which trained behaviors tend to be plastic are the presence of rewards or the absence of punishers.

constitute the many paths I might take to get a beer were that my primary goal. Many of the courses of action I would be able to take have never been rewarded or reinforced or selected for in my evolutionary history. My ancestors may have all been faithful Amish folk; I may never have purchased beer before; or I may never have done so at the gas station around the corner. Nevertheless, I can still successfully get beer by these and other means if I have seen signs in the gas station window advertising beer or if I have seen other people purchase beer at other gas stations or if I have heard people talking about the wide selection of beer on offer at the liquor store on Jefferson and 14th…etc. It is difficult to see how any simple lawlike connection between my experiences with beer (the input) and my behavior (the output) could predict the indeterminate number of paths through the world that I might take to achieve what (to me) is the same kind of outcome.[11] By contrast, when we posit an internal motivational state of desire, we attribute a state that tracks the desired outcome and inclines the agent toward it, across the manifold paths that might have lead to it (or to the multiple ways of realizing the end state). Such an explanation posits fewer entities and without ad hoc assumptions (e.g. a multitude of fixed action patterns, or reinforcement relations that depend on extreme response generalization). When behavior is sufficiently plastic (especially if the range of plasticity includes novel or unreinforced behaviors), the evidence favors the existence of motivational states that guide behavior in this highly variable way.

1.2.3 Exchangeability of behavior and the explanatory role of informational states

Whereas plasticity focuses on the various means by which an organism can achieve a given end, another condition, *exchangeability*, focuses on the various ends that

[11] The goose's behavior has some degree of variability, but this variability can be predicted by a lawlike input/output relationship. In a multidimensional perceptual space characterizing the stimulus (the "egg"), there is likely to be a single region in that space that triggers the same kind of egg retrieval response (cf. Sterelny 1999). The same cannot be said for the many behaviors that are instrumental for beer retrieval. There is no multidimensional perceptual space on which these behaviors are triggered only by stimuli in a single region and there is no principled taxonomy of behaviors that can characterize my beer retrieval behaviors as of the same kind.

an organism can pursue by similar means. Two different kinds of end state are exchangeable with respect to an element of an organism's behavioral repertoire if that element can be used to bring about either end state. If an organism is only able to use a given behavioral sequence in service of one kind of end state, then we have little or no reason to credit the organism with selecting that element of its repertoire in service of a given purpose. If the graylag goose can only use its beak to maneuver its eggs to its nest and cannot use a similar movement to accomplish some novel end (say getting food or performing a trick for a reward), then we have little reason to credit the goose with selecting this behavior to accomplish either end.

Why does exchangeability give us reason to credit an organism with structuring its behavior? Exchangeability supports hypotheses that postulate informational states representing the behavioral means available to (or the affordances of) an organism. Imagine the difficulty with which an ethological or behaviorist hypothesis might explain all the ends that I can pursue by getting in my car and driving to the gas station. I may never have won the lottery or bought beer or tried to buy something off of craigslist. Nevertheless, I am still able to drive to the gas station in order to meet up with a seller on Craigslist or to buy a lottery ticket or a beer. When we posit informational states of belief concerning an agent's abilities or available means, we attribute to its possessor the ability to deploy those means across the many ends to which those means are appropriate. When behavior is exchangeable (especially if it can be applied to novel ends), we have evidence for informational states that allow the deployment of abilities or means toward several different ends.

When both exchangeability and plasticity are present in a given means-end pairing, they demonstrate a kind of agential integration across means and ends. Susan Lackey calls this *holism*:

> The holistic flexibility of intentional agency contributes a degree of generality to the agent's skills: a given means can be transferred to a novel end, or a novel means adopted toward a given end. The end or goal functions as an intervening variable that organizes varying inputs and outputs and allows a degree of transfer across contexts. As a result, *understanding another organism as an intentional agent permits transfer or generalization from a specific circumstance/behavior contingency to others*: if she has ends that call for deception, she may be expected not only to give leopard alarm calls when there are no leopards, but also to give eagle alarm calls when there are no eagles. (Hurley et al. 2003, 237–238 emphasis mine)

This kind of holism provides strong justification for attributing the structuring of behavior to the organism (rather than to external sources). The ability to integrate information and motivation across the situations that one encounters is part of what it means to be an agent possessing motivational and informational states.

To sum up, purposive behavior is behavior that has a psychological explanation, which appeals to the motivational and informational states of an agent. When behavior is plastic and exchangeable, we have evidence that behavior is structured by motivational and informational states that interact holistically and thus are attributable to the organism performing the behavior.

1.3 When do emotions motivate purposive behavior

Now I shall address a final clarification. There are many ways that emotions influence behavior, but under what conditions do emotions *motivate* purposive behavior? To recapitulate some of the conclusions of the previous section, the point of attributing a motivational state to an organism is to explain the ability to reach the same end state across a range of different means. Moreover, to play that explanatory role, the motivational state has to incline the organism toward the same end state across several different ways of achieving it. Given this way of understanding of motivational states and

their role in producing purposive behavior, emotions function in this way only if they incline an organism toward some specific end state across different ways of achieving it. In this section, I look at several of the ways that anger influences behavior. I show that in many of these cases, anger does not actually motivate behavior because the inclination toward the end state of behavior is a motivational state that is independent of anger. The point is to clarify the distinction by drawing attention to some of the ways that anger can *seem* to motivate action without actually doing so.

First, suppose that I am angry. As a result, my body may be in an unpleasant state of increased physiological arousal (heart rate, blood pressure, etc). Psychologists have long known that aggression, whether verbal or physical, can cause a reduction in physiological arousal and that this contingency can increase the likelihood of aggression (probably by reinforcement).[12] Thus, we might suppose that being angry causes me to say something rude to a bank teller because of the change in arousal that it affords me. In an indirect sense, I was rude to him because I was angry. Nevertheless, my anger is not the motivational state responsible for this action. Rather, if my purpose is to diminish the unpleasant state of arousal that my anger produced, then the relevant state is something like a desire to diminish that arousal. To satisfy this desire, I might just as easily have told him a joke (Newman and Stone 1996) or breathing slowly, actions that we would be disinclined to attribute to my anger. Alternatively, I might have been just as rude had I just returned from a run. In such cases, I would have acted *for the very same reason*, to influence my unpleasant bodily arousal (Zillmann 1979). In this case, the end state toward which I am motivated is a state of diminished arousal, and that motive is independent of anger. I can be motivated in this way without being angry (e.g. due to

[12] See Tyson (1998)for a review. Some of the same things can be said about negative affect. See Berkowitz, Cochran, and Embree (1981) for examples.

physical activity) and the motivation can lead to actions that do not accord with anger (e.g. telling jokes or breathing deeply and slowly).

Emotions can also influence action through affective forecasting. The idea is that in considering a possible course of action, my (offline) emotional response to the anticipated outcome will help me to decide whether or not to act in that way (e.g. Baumeister et al. 2007).[13] Deliberation that is influenced by affective forecasting thus depends more specifically on the anticipated emotional states, rather than just on their effects (e.g. bodily arousal or affect). Nevertheless, this is still not an example in which anger is the motivational state responsible for action. To see this, imagine that someone has decided to turn on the radio and that she is considering whether to tune in to Rush Limbaugh (which usually makes her angry) or to tune in to the classical music station (which makes her less angry). Suppose that she is driving to work and running late because someone slashed one of her tires. Consequently, she is very angry. Will her anger make her more likely to choose one option over the other? We cannot say. Whether it does or not depends on her particular constellation of attitudes toward anger. Perhaps she enjoys being angry. Perhaps anger helps her to perform better in her competitive workplace (e.g. Tamir, Mitchell, and Gross 2008). If so, better to listen to Rush Limbaugh. Alternatively, it may be that anger has never been for her a source of empowerment or positive motivation. Perhaps she has learned to cope with these situations by turning her anger inward and this contributes to an inhibited or depressed mood (Smits and Kuppens 2005; Felsten 1996; Bridewell and Chang 1997). The fact that one option will make her more angry and the other less does not directly influence her decision. Rather the effect of anger will be mediated by the attitudes, habits and coping mechanisms that have grown up around her anger and the situations that elicit it. It

[13] One might think that Damasio's (1994) somatic marker hypothesis is relevant here. However, he emphasizes a very different role for emotion in deliberation. What distinguishes his view from affective forecasting is an emphasis on the informational role of emotions rather than their motivational role. Thus, his account does not illuminate a separate motivational role for emotions.

should be clear that anger is not the motivator in this case. Whatever radio station she might choose, the end state toward which she is motivated depends not on her state of anger, but on the cluster of desires and other attitudes she has concerning anger. If she chooses Rush Limbaugh, it is because of her motivation to increase or maintain her anger, and this motivation does not depend on her anger but instead on her attitudes toward it.

One conclusion that might be drawn from these examples is that emotion cannot motivate behavior through deliberation. However, this is not so. If motivational states like desires can influence behavior via deliberation as in these examples, then there is no obvious reason why emotional states could not do so as well, provided that they motivate agents toward an end state of some kind. Suppose for instance that anger motivates me to retaliate. That is, independent of my attitudes about anger, when angry, my behavior is guided toward end states that constitute retaliation rather than a range of other outcomes, such as reconciliation or avoidance. One could imagine this very motivation influencing a deliberate plan to avenge a past insult. If this is possible, then in such a case, the influence of anger is unmediated. It might not hold complete sway over my decision, but it would motivate me toward a specific kind of outcome (retaliation), irrespective of my attitudes or dispositions toward anger.[14] This is the sense of emotional motivating that I am interested in here. Thus, the central question of whether anger motivates behavior is the question of whether states of anger incline me toward a single kind of action irrespective of my attitudes or dispositions regarding anger. It seems to me that anger could exert this kind of influence whether or not my ultimate course of action is arrived at via deliberation.

[14] It is worth noticing that a number of seemingly distinct behaviors count as retaliation. For instance, one might retaliate toward verbal offenses with verbal attacks, whereas toward relational offenses, one might merely stomp out of a room. While the latter isn't a paradigmatic case of retaliation, it can play that role if one knows that the other person in the room can be emotionally distressed by outbursts of that kind.

Importantly, some philosophers emphasize the flexibility of emotional behavior in humans and insist that their behavioral effects depend entirely on planning and deliberation. For instance, Prinz distinguishes between motives, or reasons for action, and motivations, or impulses to act. On his view, emotions can be reasons for action, but they give rise to motivations only if the agent chooses:

> Being angry provides a reason, ceteris paribus, to attack…But emotions are not always motivations. They do not always succeed in impelling us. One can be angry, it seems, without being disposed to revenge. In contrast, one cannot be hungry without being disposed to eat. The link between emotions and action tendencies is weaker than the link between motivation and action…In the case of anger, our bodies are prepared for aggression, and the valence marker tells us that we should maintain that state (positive valence) or change that state (negative valence). At this point in processing, no action has been selected, no strategy has been determined, no plan has been conceived. The somatic state and valence marker *must be fed into a mental system that selects responses*. Among the available responses is violent revenge against the source of our anger. The state of anger increases the probability of this response, but it is not constituted by this response. The decision to seek revenge is a choice that *follows* anger. Once that choice has been made, we can say there is action tendency at work. The action tendency is not itself a motive for action. It is a motivation. An active plan to seek revenge is an urge or a want; it is like hunger." (J. J. Prinz 2004, 193–194)

The upshot is that on Prinz's view, the influence of emotion on behavior is primarily determined by deliberation or planning. Whether or not the emotion gives way to motivation depends on whether a certain motivation is *chosen* through deliberation.

Tappolett captures the general outlines of this view when she describes what she calls the desire model of emotional motivation as applied to fear:

(a) given its physiological underpinnings, fear facilitates but does not necessitate certain types of [fixed] actions;

(b) fear involves a desire that sets a goal, such as the avoidance of a specific harm or loss, and if it results in action, it does so only on the basis of the agent's deliberation. (Tappolet 2013)

Tappolet follows Prinz in claiming that emotions lead to desires, which have their behavioral effects through deliberation or planning rather than directly motivating behavior.

This view has three major drawbacks. First, this view neglects the possibility that emotions can influence deliberation.[15] For instance, it could be that if I deliberate while angry, my valuations of outcomes are systematically skewed, making vengeful outcomes seem far more favorable. In that case, it would take great effort to choose a course of action that was not vengeful, and it would be misleading or false to say (as Prinz does) that I was not disposed to revenge in a robust sense. If this scenario is an empirical possible, then anger can create a disposition toward revenge without me deliberately choosing revenge. This is a possibility that Prinz does not explicitly consider.

Second, this view leads to inaccurate predictions. If anger leads to revenge "only on the basis of deliberation", then inhibiting deliberation should reduce vengeful behavior when angry. Under the plausible assumption that depleted resources for self-control interfere with deliberation, a natural prediction (based on Tappolett and Prinz's view) is that resource depletion will decrease the likelihood of vengeful behavior when angry. The opposite is actually true. For instance, in several different studies DeWall and colleagues (2007) exposed participants to different resource depletion manipulations (abstaining from eating a donut, not looking at words that appear during a video viewing, reading color words printed in incongruent ink colors, and breaking a habitual behavior)

[15] See Lerner et al (2006) for a recent review of the systematic effects of anger on deliberation.

prior to a provocation. While resource depletion did not change reported levels of frustration or anger, they did increase several measures of aggression against the provocateur (making them eat a cracker with more hot sauce, subjecting them to higher intensity noise blasts, and giving them negative performance evaluations for a job opportunity) as compared with control participants (who were not subjected to resource depletion).

Finally, Prinz and Tappolet have to deny the possibility that emotions cause impulsive or unplanned *action* like kicking a door in anger (which I take to be distinct from mere *behaviors* like involuntary facial expressions). While there is little psychological data that would explicitly contradict this view, there is a wealth of evidence showing (as do the experiments just described) that aggressive actions require active exertion of self-control for inhibition (see Denson 2009 for a review). There is also a wealth of common sense data. Since I do not want to rule out any of these empirical possibilities (or plausible hypotheses), I will not follow Prinz and Tappolet in their claim that emotions can only motivate action through deliberation.

Thus, the question of whether anger motivates behavior depends on whether anger inclines the agent to select behaviors because they lead to a certain kind of end state irrespective of an agent's attitudes or dispositions toward anger and independently of whether she deliberates and decides to give rein to her angry impulses. Does anger ever motivate purposive behavior of this kind? I doubt that anyone capable of anger needs to consult psychological studies to answer this question. I suspect that only the most even-tempered or peacefully-ensconced have not experienced an immediate urge to inflict verbal or physical harm when angry. Who has not (during childhood if not as an adult) spontaneously struck a door, or a table or a hammer after inadvertently stubbing her toe or biting her lip or bashing her hand? In such cases, I think there is little reason to doubt that something answering to the term "anger" motivates behavior. Yet there

may be some reason to doubt whether such an emotion is the *shared emotion* of anger, or the emotion that is homologous with anger in rats. In section 3, I consider some of the reasons that one might doubt this.

2. Resolving the Paradox

Now that some key distinctions have been made, I restate the paradox:

1) Some shared emotions are continuous between human and nonhuman animals.
2) In humans, these shared emotions motivate purposive behavior.
3) In animals, these shared emotions only cause stereotyped behaviors, which are not purposive.

In this section, I demonstrate that 3) is false concerning anger in rats. In the following section, I develop some of the implications of this claim for emotion theory.

2.1 Initial thoughts

Given discussion in section 1.2 and in chapter 2, there are initially several reasons to think that angry behaviors in rats would be non-purposive. If the selection model of chapter 2 is correct, a great deal of the structure of the behavior is shaped by natural selection. First, the nonlethal nature of a resident rat's attacks (constituted by the fact that they are directed at a protected target site) is probably structured by kin selection. Second, consider the resident rat's tendency to attack an unfamiliar male intruder independently of whether these attacks have ever met with success. If the arguments of chapter 2 are correct, then this behavioral tendency was structured by the demands of frequency dependent selection.

Were we to inquire as to the structuring causes of the resident rat's behavior, we would ask a question like the following: why does an unfamiliar rat intruding on a resident's territory *cause the resident rat to* bite the intruder's back until it runs away? The answer is that rats with heritable dispositions to behave in these ways were more

likely to promote copies of the same heritable dispositions in subsequent generations. So the structure of these behavioral dispositions was determined by selection and not by a structuring cause internal to individual rats, such as their informational and motivational states. It would be absurd to think that the rat has any kind of insight into the outcomes that tend to result from these behaviors. More specifically, we have no reason to attribute to the rat the purpose of running off a competitor or preventing the propagation of mutant strategies nor do we have any reason to credit it with choosing the manner in which these goals are pursued (by biting the back and thus avoiding lethal harm to intruders).

Another reason to doubt that these behaviors are purposive is their *apparent* lack of exchangeability and plasticity. The biologists who study these attack behaviors are quick to say that they are highly stereotyped, meaning that the same patterns of behavior can be observed in any appropriately stimulated resident rat. In other words, resident rats seem to have only a few behavioral means by which to pursue the end of biting the intruder's back. Neither do these behavioral means seem to be available to the rat for deployment toward other ends. Thus, it is initially doubtful that a resident rat chooses the means by which this end is achieved; doubtful that the rat has any informational state representing those means across the ends to which it might direct them; and doubtful that the rat has a motivational state that explains its ability to bite the backs of intruders across a range of means to that end.

2.2 Evidence for plasticity and exchangeability

Nevertheless, some of these appearances are misleading, and this becomes apparent when we consider experimental investigations of these behaviors. In certain conditions, the presence of the unfamiliar male can produce highly flexible and novel behaviors that are clearly aimed at biting the intruder's back. If an intruder rat is tied down on a Plexiglas plate with only its ventral surfaces (belly-side) exposed and placed in

the cage of a resident rat, the resident will sometimes bite at the bands that tie down the intruder or dig under the intruder so that the resident can bite the intruder's back (R. J. Blanchard et al. 1977). In contrast, none of these behaviors are adopted when the intruder is tied down with his dorsal surface exposed.

These behaviors are clearly not stereotyped forms of attack, rather they are forms of flexible behavior adjustment to achieve the aim of biting the intruder's back: they exhibit both plasticity and exchangeability. The behavior is plastic because the same end state can be achieved by several, novel means. Attempts to bite the intruder's bonds or to dig underneath the intruder are novel means toward the end of biting the back of the intruder. This suggests that the rat has a motivational state by which it can arrive at the normal end state of behavior (biting the intruder's back) by various means. Since the rat can reach the end state through novel routes, there are no reasonable hypotheses concerning the rat's behavior that could predict these routes without appealing to a motivational state of this kind.

Moreover, some of a resident's behaviors are exchangeable because the same means can be deployed toward different ends. Digging is an element of the rat's behavioral repertoire that is used for an entirely different purpose: constructing burrow systems for shelter and nesting (Boice 1977). The rat can thus exchange digging as a means for different ends. This suggests that the resident has informational states (representing its available means) by which it can deploy digging behaviors toward different ends. Since both of these conditions, plasticity and exchangeability, are satisfied concerning the same end and for some of the means by which it is pursued (respectively), there is some reason to believe that the motivational and informational states of the rat can be integrated across different contexts. There is a many-to-one and one-to-many mapping from a resident rat's informational states (representing its available means) to the motivational states with which they can interact. Thus, we have considerable reason

to think that biting the intruder's back is the resident's own purpose in acting and reason to credit the resident with an ability to select the means by which this end is pursued.

Importantly, the emotional state of anger in the rat seems to motivate this behavior. The innate (or at least highly invariant) disposition to bite the backs of intruders is coordinated with the cluster of behaviors and physiological changes that enable it to defend its territory. This is why it is included in the set of phenomena explained with reference to the underlying psychological entity, anger. Moreover, the innate disposition to bite intruders' backs seems to be responsible for the novel behaviors manifested in the bound-intruder task. The novel behaviors are structured to bring about the same end state as the stereotyped behaviors that usually bring it about, and the most compelling explanation is that the same underlying motivational state is responsible for inclining the rat toward both stereotyped and novel behaviors. In sum, anger in rats motivates purposive behavior.[16]

What is unique and important about this singular example is that it is a rare case in which instrumental behaviors are clearly connected with a well-characterized emotion system. There is a wealth of anecdotes concerning animal behavior that suggest the phenomenon captured in this experiment is not atypical. Some of these anecdotes concern primates. For instance, ethologist Marc Bekoff (Bekoff 2009, 81) tells a story about a driver in Saudi Arabia who hit and killed a baboon.

> Afterward, the baboon's troop lay in waiting for three days by the side of the road until the same driver appeared again. As the driver passed the troop, one baboon screamed and then all the baboons threw stones at the car and tore out its windshield.

Obviously, the behavior described was highly flexible, and the details of the case suggest that something akin to payback was the motivation behind the behavior.

[16] On my view of psychological explanation, this is consistent with the claims above that resident rats do not have motivational states representing other outcomes of the behavior (e.g. running off an intruder or preventing the propagation of mutant strategies). The experiment does not provide any evidence that the rat's behavior is plastic or exchangeable with respect to these goals.

A similar anecdote involves a Siberian tiger (Vaillant 2010). After Vladamir Markov wounded the tiger and took a portion of the tiger's kill, the tiger found Markov's hunting cabin, destroyed many of Markov's possessions and lied in wait for him for 12-48 hours. When Markov returned, the tiger killed and ate him. Hunger may have partially motivated the tiger's actions. Nevertheless, at least part of the motivation here seems to have been payback for the wounding or theft. Otherwise, it is difficult to explain why the tiger would have also destroyed Markov's possessions.

In a more scientific report on aggression induced by electrical brain stimulation, Jose Delgado described aggression in six macaques:

> The aggressive behavior…was well organized, skillfully performed and specifically oriented toward the investigator or toward determined animals… The aggressive intent was reliable, but the motor pattern varied according to the proximity and reactions of the other animals. For example, monkey Charley…turned to the right or left depending on the location of monkeys #3 and #4, adjusting his speed to that of the chased animals, and moving his body and hands in order to hit or grab the flying [sic] monkey… These facts demonstrated that brain stimulation had not activated a stereotyped response or a kinetic formula, but had produced a basic change in the emotional tuning for the processing of sensory inputs. (Delgado 1967, 179)

In this example, the aggression was apparently affective in nature (though there is little indication of what kind of affect), and the behavior seems highly plastic.

These studies and anecdotes demonstrate different forms of flexible behavior that seem to be motivated by emotions like anger. They give us some reason to doubt that the emotional behaviors of rats are atypical. However, by comparison with the rat experiment, these other examples lack as clear a connection with a specific emotion system, and it is less clear what the aim of the attacks are. The experimental design and the connection with back biting are what make the rat example a particularly clear case

for my purposes. Nevertheless, the other examples suggest that the singularity of the example is no indication that it is atypical or anomalous with respect to aggression and other emotional behaviors in the animal kingdom.

3. Implications

Of course, one response to this argument will be that it is trivial or obvious that emotions motivate purposive behavior in nonhuman animals. However, the argument has nontrivial implications for a view held by at least two influential emotion theorists, one that has modest theoretical support: even in humans, shared emotions like anger do not motivate purposive behavior.

One primary reason to take this position concerns the role of basic emotions in explaining various phenomena. Basic emotions in humans are primarily postulated to explain highly stereotyped behaviors such as involuntary facial expressions of emotion. Thus, there is some hesitation to lump purposive behaviors together with the other behavioral phenomena that basic emotions explain. Consider Paul Ekman's (1977) view:

> Somewhat longer and more elaborated [than facial and physiological responses produced by basic emotions] are the coping behaviors directed at whatever has set off the emotion. Included would be fighting, fleeing, denying, apologizing, etc…
>
> Through experience, with sufficient time and learning, habits become established for how to cope with each emotion. I do not believe that such coping behaviors are part of the *given* [basic emotion]…Memories, images, expectations associated with one or another emotion are, like coping, *not given but acquired*… (Ekman 1977, 56–57 emphasis mine)

For Ekman, what is important is that the *given* or innate (and not acquired) basic emotion is supposed to explain pan-cultural response tendencies toward emotion elicitors. This seems to be one of his reasons for thinking that basic emotions do not include coping behaviors, since these behaviors seem to be acquired.

Neither does Ekman think that basic emotions are capable of strongly influencing the acquisition of coping behaviors:

> Biology may provide some predispositions affecting the likelihood of one versus another type of coping behaviour being developed for an emotion. For example, the skeletal muscle response for anger suggests that attack may become more frequent for coping with anger than flight. Yet, this predisposition is relatively fragile. Experience can overcome such predispositions and institute diametrically opposite coping. Coping involves a wide range of elaborated activities, and biology at best gives only a tap in a direction. Culturally and individually variable learning is the overwhelming contributor to coping. (Ekman 1977, 64)

So far as I know, Ekman retains this view up to the present. Ekman (2003) writes the following in an endnote: "Frijda's description of the actions that characterize each emotion includes what I have said and quite a bit more. I believe it is only these rudimentary initial postural moves [e.g. looking down on an object of contempt, fixed attention on the object of surprise, movement toward the source of sensory pleasure, and slumping posture and loss of muscle tone in sadness] that are inbuilt, automatic, and universal." (p. 268)[17] I suspect that one of the motivations for this view is a hesitation to

[17] Contrary to the way Ekman and others (e.g. Clore 1994) frame the matter, the question is not whether anger includes inbuilt, automatic action patterns, as if a jab to the nose or some other action were written into human DNA. The question concerns whether anger can motivate a person toward a specific, biologically predisposed end state (say retaliation) by inclining an agent toward a range of (and possibly acquired) means, including a jab to the nose, a withering glare, or a variety of other possibilities. Ekman gives entirely separate reasons to doubt this possibility:

> Compare with coping the initial skeletal muscle response directed by the [basic emotion] when an anger elicitor has been identified…The immediate skeletal muscle response might be a slight movement forward. *Coping could vary* – attack, flight, denial, appeasement, etc. We discover how to cope with our emotions, what is likely to be successful, proper or improper. When angry, our likelihood of fighting or scratching our face, depends upon what we have learned about how to deal with the particular kind of anger elicitor.
>
> Once coping techniques have been acquired, they can become so well learned that they operate automatically and are called forth when the [basic emotion] is set off… (Ekman 1977, p. 72)

The contrast here is between the incongruity of coping behaviors (perhaps mediated by learning) and the specificity of purposive behaviors. Ekman's idea is that if anger was accompanied by a motivation toward a specific end state, then it would not "call forth" so many different and seemingly incompatible coping

extend the explanatory role of basic emotions to the control of flexible or acquired behaviors.

The problem here is that this theoretical consideration is frail in the face of empirical evidence to the contrary. If basic emotions could only predispose or motivate behaviors that are inbuilt, automatic and universal, then we should be utterly surprised to find that anger in rats can motivate novel back-biting behaviors. Perhaps digging and biting are innate, universal and at times automatic, but they are certainly not innate, universal, or automatic responses *to an intruding conspecific*. Moreover, the fact that anger can motivate such behavior in rats should make one extremely dubious of any purely theoretical rationale for claiming that human anger cannot also motivate action (as opposed to mere stereotyped *behaviors*).

I suspect that Ekman has another reason to think that basic emotions do not influence purposive behaviors. This arises from a focus on their signaling function. For instance, he thinks that "…the primary function of emotion is to mobilize the organism to deal quickly with important *interpersonal* encounters, prepared to do so by what types of activity have been adaptive in the past." (Ekman 1999) The anchoring of basic emotions to facial expressions is one of the reasons to think they have a central signaling function. Basic emotions aid in the avoidance of poisons, parasites and predators and help to deal with gains, losses and resource competition, and for groups of organisms (whether closely related or highly interdependent) to effectively deal with each of these tasks,

---

behaviors. Anger would almost never be accompanied by behaviors like flight, denial, and appeasement if it included a biological predisposition toward aggression or retaliation.

I suspect that Ekman is assuming here that it is anger, rather than some other psychological entity, that automatically calls forth many of the seemingly incompatible coping behaviors. This assumption is entirely unwarranted, and there is some support for the opposite assumption. For instance, there is some evidence that anger regulation can occur automatically (Mauss, Cook, and Gross 2007; see also Smits and Kuppens 2005). If this is right, then the fact that behaviors like flight, denial and appeasement can occur subsequent to anger is consistent with the claim that anger includes an opposing action-tendency toward, say, confrontation. Given the existence of automatic emotion regulation, one cannot infer the non-existence of action-tendencies from the fact that they are not consistently manifested in behavior after an emotion has been elicited.

signals are critical. A look of disgust can warn others that food is contaminated; an alarm call can alert others to a predator nearby; and a threat display can signal a willingness to escalate a competitive encounter (J Archer and Huntingford 1994). In effect, humans and other animals avoid poisons, parasites and predators together, and they negotiate resource competition primarily with other members of their species (Dawkins 2006, chap. 5).

If this is right, it makes sense that basic emotions are innate or at least develop with a high degree of regularity across a wide range of environmental conditions. Consider Ernst Mayr's view: "Since much of the behavior directed toward other conspecific individuals consists of formal signals and of appropriate responses to signals, and since there is a high selective premium for these signals to be unmistakable, the essential components of the phenotype of such signals must show low variability and must be largely controlled genetically." (Mayr 1974, 657) Signaling phenomena of this kind will thus tend to be structured by selection processes and strongly influenced by inheritance rather than by internal motivational and informational states of animals. Thus, if basic emotions are postulated as proximate explanations for signals of this kind, this may be another source of hesitation to extend their explanatory role beyond innate behavioral dispositions.

Nevertheless, the flexibility of a resident rat's behavior overturns this theoretical rational as well. Just as with signaling phenomena, we have good reason to believe that *parts* of this behavior program are innate or highly constrained by genetics (cf. section 2.2). Nevertheless, this clearly does not warrant the conclusion that the program will *only* produce genetically constrained behaviors. It is highly unlikely that the resident rat's digging underneath the intruder and biting at the intruder's bonds are genetically constrained in any interesting sense. Rather, this example suggests that evolutionary

forces can structure one or more aspects of the action (e.g. the goal of back biting plus some of the means for achieving it) while leaving room for improvisation.

Another emotion theorist, Paul Griffiths gives similar theoretical reasons to think that the behavioral influence of basic emotions is limited to stereotyped behaviors. Griffiths argues that basic emotions cannot, without significant revision, replace folk concepts of emotion. This argument depends critically on the theoretical role of basic emotions: "The identification of emotion in general with [basic emotions] would exclude a lot of what is currently regarded as emotion from the revised category… it would be argued that [many of the psychological states referred to by folk emotion concepts] are *too flexible, too well integrated with long-term, planned action*, and so forth. The extension of the emotion concept would be restricted to short-term, *stereotyped responses*…" (P. E. Griffiths 1997, 1997:241 emphasis mine)[18] On this view, anything answering to the term "anger" that is integrated with long term planned action is unlikely to be the same kind of psychological state as basic human anger. Griffiths compares this reasoning to Ekman's (1985) argument that the startle response is not an basic emotion. While it does have a characteristic facial expression, Ekman thinks the startle response is too reflexive and too difficult to suppress to count as a basic emotion. Likewise, Griffiths thinks that some folk emotion concepts should be excluded from the category because they refer to psychological states that are too flexible. Thus, Griffiths seems to think that the explanatory role of basic emotions should be secluded to highly stereotyped behaviors.

But how flexible is *too* flexible on Griffith's account? Griffiths argues that basic emotions are inflexible in the sense that they lack integration with certain cognitive processes. These are "…the processes in which people use the information of the sort

[18] I think it remains plausible even in light of the conclusions of this essay that affect program states would exclude a lot of vernacular emotions. There are plenty of emotional states that are not shared across cultures and are not plausible influenced by the phylogenetic information encoded in affect programs.

they verbally assent to (traditional beliefs) and the goals they can be brought to recognize (traditional desires) to guide relatively long-term action and to solve theoretical problems." (P. E. Griffiths 1997) Griffiths thinks that the involvement of basic emotions with these processes is limited in several different ways (P. E. Griffiths 1997, 1997:100). Basic emotions unfold automatically and involuntarily without requiring elicitation or guidance from higher cognitive processes. They are opaque to cognitive processes: "People are aware of [basic emotion] outputs, which are the emotional responses themselves, but not aware of the processes that lead to them…" Finally, they are informationally encapsulated:

> [They] cannot access all the information stored in other cognitive systems, and [they] can store information that contradicts that other information. Conscious beliefs concerning, for example, the harmlessness of earthworms do not get taken into account when the system is deciding upon a response. (P. E. Griffiths 1997, 1997:93)

In other words, basic emotions are not holistically integrated with the conscious beliefs and intentions that guide much of human action. In the end, Griffiths comes short of denying that emotional behaviors lack holistic integration with *any* of the representational states of an organism. Nevertheless, he only considers two possible influences that emotions might have on behavior. One is that they trigger stereotyped behaviors, and the other is that they are integrated with beliefs and desires and so also with long term planning. This omits the possibility that emotions *are* highly integrated with representational states of organism *aside from conscious beliefs and desires* and the possibility that they might do this while also causing stereotyped behaviors.[19]

Moreover, emotions may cause action in both human and nonhuman animals in just this way. The plasticity and exchangeability of the resident rat's back biting attack

[19] Though his recent work (P. E. Griffiths and Scarantino 2004; P. E. Griffiths 2010) comes closer to admitting this kind of possibility.

could be produced by something far less cognitively complex than the desires and beliefs that seem to guide long term planning in humans (perhaps an urge to bite the back coupled with a representation of a "digging-for-biting" affordance). Similarly, when I kick my car door in anger, the action seems to be produced by informational and motivational states less complex than conscious desires and beliefs (cf. Ginet 1990). In my experience, such actions occur without any forethought or planning, and it is difficult to think of any conscious belief or desire that I had prior to acting or that I would report to explain my action. What desire or intention would plausibly cause me to kick the door? It is not likely that I want the door be kicked or harmed, and it is difficult to think of a desire that fits the bill. Perhaps I just wanted to express my anger, but I cannot recall ever experiencing a conscious desire of this kind.

On the other hand, if the anger of a resident rat motivates purposive behavior because *it is* sufficiently integrated with its beliefs and desires, then the role of beliefs would not be restricted to long term planning and the requirements for integration would seem to be minimal. We could easily assume that human anger is just as integrated with beliefs and desires.[20] Either way, the lack of integration with beliefs, desires and long term planning processes does not give us any strong reason to doubt that basic emotions motivate action. Either anger in rats motivates action without being so integrated, or, if anger in rats motivates action *because it is* integrated with beliefs and desires, then there is little reason to doubt that humans share the requisite integration for anger to motivate action. In sum, these theoretical considerations seem frail when juxtaposed with the demonstrable flexibility of animal emotions.

4. Conclusion

[20] Notice that this would require modification of Griffith's view that beliefs and desires are reportable or consciously accessible.

Moreover, some of the evidence surrounding human aggression seems to provide some evidence favoring a shared anger motivation in humans. Not only is there strong evidence that both anger and physical aggression appear very early in human development (Côté et al. 2006), there is some evidence that flexible aggression in young children is coordinated with the other components of the anger response. Hubbard et al (2010) report that in a pilot study angry facial expressions became synchronized with skin conductance responses prior to acts of reactive aggression (disfiguring a virtual interactor's artwork after receiving criticism) in young children but not prior to acts of proactive aggression (disfiguring artwork when a reward could be obtained).[21] The kind of aggression that follows these expressions is not inflexible or stereotyped, rather it requires integration with knowledge of how another child would respond to having their artwork disfigured.

Of course, a key contrast between human aggression and rodent aggression is that anger in rats has the fixed aim of biting the back (though see Carrier and Morgan 2014 for some evidence that the face has been a primary target of hominin aggression). While disfiguring another child's artwork might qualify as back-biting in a metaphorical sense, it is not easy to imagine that it is motivated by the same kind of state that motivates rats to engage in literal back-biting. Moreover, it is possible that over the long course of evolution from our common ancestors with rats until today, the influence of shared anger on instrumental behavior was simply bred out of our ancestors, and that the impulse toward reactive aggression is acquired developmentally.

Nevertheless, a theoretical case can be made that rather than disappearing, shared anger became more flexible in the evolution of the primate and then hominin

[21] In general, anger is closely associated with the reactive subtype of aggression but not the proactive, predatory, and instrumental subtypes of aggression (with which reactive aggression is usually contrasted, see Vitiello and Stoff 1997 for a review). For instance, proactive and predatory forms of aggression (even in other species) need not involve anger and are often unaccompanied by the distinctive physiology of anger.

lineages. In ethological comparisons of rats and macaques, there is already a notable change in the dynamics of aggressive interactions between dominant and subordinate macaques (D. B. Adams 1981b; D. B. Adams and Schoel 1982). While back-biting is preserved, there are no fixed strategies for offense or defense in the macaque. This may reflect increasing degrees of freedom for aggressive interaction, degrees of freedom that only increase as social cognition becomes more complex.

Here is the trend as I understand it. In tandem with increasing social complexity, resource competition becomes increasingly abstract. Position in a social dominance hierarchy becomes the main determinant of reproductive success, so defending a position in a hierarchy takes the place of defending a physical territory. Even in rats, when population density increases, the mating system shifts from polygynous (where the alpha rat monopolizes estrus females) to polygynandrous (where males copulate sequentially with estrus females). Since dominant rats in the latter mating system have reproductive priority and probably greater success, defending a position of dominance from cohabiting (and potentially rivalrous) males becomes as important for reproductive success as defending the colony from unfamiliar males.[22] Accordingly, the dominant rat in a colony (with a mix of male and female rats) reinforces dominance by the same patterns of confrontational aggression exhibited by resident rats in experimental settings.

Nevertheless, as social cognition increases in complexity by comparison, a broad range of possibilities arise for defending position in a dominance hierarchy with minimal energy expenditure and minimal risk of injury. Instead of repeatedly biting the back of a subordinate conspecific, organisms can respond to challenges and reinforce dominance with a threatening facial expression or by physically displacing a subordinate with impunity. In humans, the abilities to wield symbols and language make it possible to

[22]Though there is rarely anything resembling an orded hierarchy except that one male, the alpha, initiates and wins most fights.

"put someone in their place" with symbolic or verbal barbs and blows. Certainly, there is the possibility of physical escalation, but it is advantageous even for an unassailable individual to waste as little energy as possible "putting someone in their place". The increase in available options for defending social dominance and its extensions (perhaps deference or respect) suggests that the motivational aim of anger would gradually become more diffuse, while still having a kind of central direction. This is roughly the path by which a motivation with a fixed aim could gradually come to resemble a vengeance motive.

In this chapter, I began by articulating a paradox. It is implausible to claim that a shared emotion like anger is continuous between humans and other animals and that it it motivates purposive behavior in humans but only nonpurposive behavior in animals. After clarifying some central terms in the paradox, I argued that we should reject the claim that shared emotions only cause nonpurposive behavior in animals. Finally, I drew out some substantive implications of this argument for emotion theory. I have concluded by gesturing at why we could expect the motivational aim of a shared emotion to become more diffuse over the course of human evolution rather than simply disappearing.

Conclusion

> I believe very strongly…that no one could ever deserve to suffer. Of the people in whose moral judgment I have the most confidence, some disagree. When some wrong-doers suffer, these people believe, this suffering is in itself good, or at least not in itself bad. Though this belief seems to me mistaken, I would be greatly relieved if I could explain why these people are making this mistake. This may be one of the cases in which an evolutionary explanation *helps to undermine what it explains*. This retributive belief may seem to justify certain natural reactive attitudes, *such as an angry desire to hurt* or the withdrawal of good will. *These attitudes are like some simpler emotions that are had by the animals that are most like us. If evolution can explain why many people have these reactive attitudes, that might give some support to the view that these attitudes, and the widely held belief that such attitudes are justified, are not responses to reasons.*
>
> -Derek Parfit (2011), p.429, emphasis mine

About halfway through this dissertation, I came across this piece of text. It has slowly dawned on me that this dissertation is, in some sense, a footnote to what Parfit says there. My main aim was to demonstrate that human anger is importantly connected to a psychological category in non-human animals, and that it is connected in such a way as to draw retributive motives under the same explanatory umbrella as some motives of non-human animals. Once we understand the structure of their shared explanation – that these motives were shaped by natural selection (and here we can add, maintained) for their biological consequences – we see that they cannot be good indicators of non-derivative reasons for punishment, such as the belief that the suffering of wrong-doers is deserved. Here, I will briefly sketch this argument once again in a revised form and point out along the way some of the work that remains to be done.

Consider again principle R: The value (or justification) of an act of punishment is not (or not only) derived from the consequences of punishment.[1] While I never gave an explicit argument, it should be clear why retributive motives incline us to believe R. Retributive motives cause people to find punishment fitting (in some sense) as a response to past wrongdoing. This explains why retributive motives make people willing to act and judge in favor of punishment (or to punish in a certain way or with a certain severity) when it will obtain no future benefit (or less future benefit than an alternative punishment). Moreover, if one endorses this kind of judgment across many individual cases, then this kind of tendency will seem to provide support to the more general principle, R.

So what does the evolutionary explanation look like for these inclinations? In chapter 2, I argued that some retributive motives are best explained by the evolutionary forces captured in the war of attrition model. For my purposes here, this means that these motives were selected for their role in preventing mutant strategies from invading the population. More specifically, I tried to show that an aggression system in rats was shaped by these evolutionary forces. In chapter 3, I then argued that human anger derives from a common ancestral trait with this aggression system. Given their connection with anger (since anger derives from a common ancestral trait), it looks plausible that retributive motives in humans and rats may have a common evolutionary explanation.

However, the question remains whether the retributive motive was maintained in the human lineage. Chapter 4 was meant to address some of the reasons that one might deny that the retributive motives of rats and those of human beings have evolutionary continuity. If the argument there is correct, then retributive motives in rats cause

[1] This principle is closely related to what Parfit says above, because the truth of a principle like R is often supposed to be made true by the fact that suffering can be deserved by trangressors and that the aim of punishment is to give transgressors their deserved suffering.

purposive behavior just as they do in human beings. Together with the fact that anger exists in both humans and rats and is connected with retributive motives in both, this gives us some reason to believe that these retributive motives have a common evolutionary explanation (for their existence in our common ancestor if not for their maintenance to the present).

If the arguments of chapter 1 are right, these motives cannot be good indicators of non-derivative value, because they were shaped by natural selection to bring about certain biological consequences. So retributive motives do not really provide evidential support for R. Nevertheless, there are two small wrinkles here that need to be smoothed out. First, principle R concerns both personal and impartial punishment, whereas the explanation of retributive motives that I give in chapter 2 (which is supposed to replace the preliminary explanation given in chapter 1), concerns only personal punishment. While there is some reason to believe that both personal and impartial punishment are influenced by some of the same underlying causes (as I argue in the introduction), it remains to be seen exactly what these underlying causes are. My hypothesis was that the repeated manifestation of retributive motives in cases of personal punishment lead to the development of a response-dependent category, the outrageous, that sustains impartial punishment judgments and behaviors. If this is correct, then we can explain impartial punishment inclinations (as well as their support for principle R) by appeal to the evolution of retributive motives in personal punishment. Thus, we have an evolutionary explanation of the broader set of inclinations (including inclinations to punish in the personal and impartial case).

Second, as it stands, I cannot entirely exclude the possibility that though retributive motives were selected for their consequences they were somehow exapted by cognitive processes of reasoning and reflection and given an epistemically virtuous connection to non-derivative sources of value they seem to indicate. Though such an

argument remains possible, I am completely unsure of how such an argument would go. It seems to me that the burden of proof is on the supporters of principles like R. Nevertheless, one way to mitigate this concern would be to give a more detailed evolutionary explanation of retributive motives in humans that fleshes out their more recent history. Some of these details involve the evolution of these motives in our common ancestors with primates, and that story was sketched briefly (though not supported by evidence) in chapter 4. Other details involve their evolution in the hominin lineage. If, as some have argued (e.g. S. Bowles and Gintis 2011), retributive motives were selected for their role in sustaining cooperation in large, human cultural groups, then we have some reason to believe that they were preserved to the present primarily because of that adaptive function. This is an area on which further work needs to be done.

If I draw on a promissory note on this last point, the overarching evolutionary story looks like the following. Retributive motives took shape through natural selection for their role in preventing the invasion of mutant strategies for resource competition. They were shaped for a subsequent role in preserving position in a dominance hierarchy (as suggested at the end of chapter 3). Finally, they were exapted for their role in stabilizing cooperation in large cultural groups. If this is their evolutionary history and if they are responsible for judgments and inclinations that support R, then we have good reason to doubt the evidential role of those inclinations. They are not good indicators that punishment has non-derivative value.

If this is right, then it may have large scale implications for western legal theory. If for instance, retributive punishment as instituted in the United States does not have good consequences overall, then we have added reason to rethink institutions of punishment.

It may also provide fuel for the fire of peace and reconciliation movements. At first glance (and perhaps due to our retributive inclinations), it is shocking that there

were no charges of crimes against humanity in the transition from Apartheid in South Africa. Nevertheless, part of the success of the peace and reconciliation movement there was due to the fact that parties to reconciliation did not seek retribution for past crimes. The debunking argument I pursue here can add some moral legitimacy for abstaining from retribution, because it can help us to see that retributive motives do not always have moral worth. While retributive motives function to deter, they can also cause reverberating waves of violence or resentment as the force of these motives are felt in response to each prior act of retribution or retaliation. In such cases, they may retain a deterrent function (say as an evolutionarily stable strategy that prevents the invasion of other strategies), but I doubt that they have any morally valuable consequences. If we can begin to see the difference, there may be greater hope for peace and reconciliation.

References

Adams, David B. 1981a. “Motor Patterns and Motivational Systems of Social Behavior in Male Rats and Stumptail macaques—Are They Homologous.” *Aggressive Behavior*.

———. 1981b. “Motivational Systems of Social Behavior in Male Rats and Monkeys□: Are They Homologous?” *Aggressive Behavior* 7: 5–18.

———. 2006. “Brain Mechanisms of Aggressive Behavior: An Updated Review.” *Neuroscience and Biobehavioral Reviews* 30 (3) (January): 304–18. doi:10.1016/j.neubiorev.2005.09.004.

Adams, David B., and W. Michael Schoel. 1982. “A Statistical Analysis of the Social Behavior of the Male Stumptail Macaque (Macaca Arctoides).” *American Journal of Primatology*.

Adams, DB. 1976. “The Relation of Scent-Marking, Olfactory Investigation, and Specific Postures in the Isolation-Induced Fighting of Rats.” *Behaviour* 56 (3): 286–297.

Albert, D J, M L Walsh, and R H Jonik. 1994. “Aggression in Humans: What Is Its Biological Foundation?” *Neuroscience & Biobehavioral Reviews*.

Albert, DJ, ML Walsh, C Zalys, and EM Dyson. 1987. “Maternal Aggression and Intermale Social Aggression: A Behavioral Comparison.” *Behavioural Processes* 14: 267–275.

Alberts, J R, and B G Galef. 1973. “Olfactory Cues and Movement: Stimuli Mediating Intraspecific Aggression in the Wild Norway Rat.” *Journal of Comparative and Physiological Psychology* 85 (2) (November): 233–42.

Alberts, JR R, BG Galef Jr, and B G Galef. 1973. “Olfactory Cues and Movement: Stimuli Mediating Intraspecific Aggression in the Wild Norway Rat.” *Journal of Comparative and Physiological Psychology* 85 (2) (November): 233–42.

Alexander, M., and A. A. Perachio. 1973. “The Influence of Target Sex and Dominance on Evoked Attack in Rhesus Monkeys.” *American Journal of Physical Anthropology* 38: 543–548.

Alexander, RD. 1962. “The Role of Behavioral Study in Cricket Classification.” *Systematic Biology* 11 (2): 53–72.

Archer, J, and F Huntingford. 1994. “Game Theory Models and Escalation of Animal Fights.” *… and Social Processes in Dyads and ….*

Archer, John. 2006. “Testosterone and Human Aggression: An Evaluation of the Challenge Hypothesis.” *Neuroscience and Biobehavioral Reviews* 30 (3) (January): 319–45. doi:10.1016/j.neubiorev.2004.12.007.

Assis, Leandro C. S., and Ingo Brigandt. 2009. “Homology: Homeostatic Property Cluster Kinds in Systematics and Evolution.” *Evolutionary Biology* 36 (2) (March 19): 248–255. doi:10.1007/s11692-009-9054-y.

Aureli, Filippo, R Cozzolino, C Cordischi, and S Scucchi. 1992. “Kin-Oriented Redirection among Japanese Macaques: An Expression of a Revenge System?” *Animal Behaviour*: 283–291.

Averill, JR. 1983. “Studies on Anger and Aggression.” *American Psychologist*.

Axelrod, Robert M. 1984. *The Evolution of Cooperation*. Basic Books.

Balleine, B W, and a Dickinson. 1998. “Goal-Directed Instrumental Action: Contingency and Incentive Learning and Their Cortical Substrates.” *Neuropharmacology* 37 (4-5): 407–19.

Bandura, Albert. 1973. *Aggression: A Social Learning Analysis*. Prentice Hall PTR.

Barash, DP, and JE Lipton. 2011. *Payback: Why We Retaliate, Redirect Aggression, and Take Revenge*.

Barnett, SA, and RC Stoddart. 1969. “Effects of Breeding in Captivity on Conflict among Wild Rats.” *Journal of Mammalogy* 50 (2): 321–325.

Baron, Jonathan, and Ilana Ritov. 2009. “The Role of Probability of Detection in Judgments of Punishment.” *SSRN Electronic Journal* 1 (2): 553–590. doi:10.2139/ssrn.1463415.

Baron, Robert A. 1971. “Magnitude of Victim’s Pain Cues and Level of Prior Anger Arousal as Determinants of Adult Aggressive Behavior.” *Journal of Personality and Social Psychology* 17 (3): 236–243. doi:10.1037/h0030595.

Barrett, Lisa Feldman. 2006. “Are Emotions Natural Kinds?” *Perspectives on Psychological Science* 1 (1) (March): 28–58. doi:10.1111/j.1745-6916.2006.00003.x.

Batson, C Daniel, Christopher L Kennedy, Lesley-anne Nord, E L Stocks, D Yani A Fleming, Christian M Marzette, David A Lishner, Robin E Hayes, Leah M Kolchinsky, and Tricia Zerger. 2007. “Anger at Unfairness□: Is It Moral Outrage□?” *European Journal of Social Psychology* 1285 (May): 1272–1285. doi:10.1002/ejsp.

Batson, C. Daniel, Mary C. Chao, and Jeffery M. Givens. 2009. “Pursuing Moral Outrage: Anger at Torture.” *Journal of Experimental Social Psychology* 45: 155–160. doi:10.1016/j.jesp.2008.07.017.

Baumard, Nicolas, Jean-Baptiste André, and Dan Sperber. 2013. “A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice.” *Behavioral and Brain Sciences* 36 (01) (February 1): 59–78. doi:10.1017/S0140525X11002202.

Baumeister, Roy F, Kathleen D Vohs, C Nathan DeWall, and Liqing Zhang. 2007. "How Emotion Shapes Behavior: Feedback, Anticipation, and Reflection, rather than Direct Causation." *Personality and Social Psychology Review□: An Official Journal of the Society for Personality and Social Psychology, Inc* 11 (2) (May): 167–203. doi:10.1177/1088868307301033.

Bedau, Hugo Adam, and Erin Kelly. 2003. "Punishment" (June 13).

Beer, CG. 1984. "Homology, Analogy, and Ethology." In *Human Development*.

Bekoff, Marc. 2009. *The Emotional Lives of Animals: A Leading Scientist Explores Animal Joy, Sorrow, and Empathy--and Why They Matter*. New World Library.

Bergquist, E. H. 1970. "Output Pathways of Hypothalamic Mechanisms for Sexual, Aggressive and Other Motivated Behaviors in Oppossum." *Journal of Comparative and Physiology and Psychology* 70: 389–398.

Berker, Selim. 2009. "The Normative Insignificance of Neuroscience." *Philosophy & Public Affairs* 37 (4) (September): 293–329. doi:10.1111/j.1088-4963.2009.01164.x.

Berkowitz, Leonard. 2012a. "A Different View of Anger: The Cognitive-Neoassociation Conception of the Relation of Anger to Aggression." *Aggressive Behavior* 38 (4) (July 13): 322–333. doi:10.1002/ab.21432.

———. 2012b. "A Different View of Anger: The Cognitive-Neoassociation Conception of the Relation of Anger to Aggression." *Aggressive Behavior* 38 (4) (July 13): 322–333. doi:10.1002/ab.21432.

Berkowitz, Leonard, S T Cochran, and M C Embree. 1981. "Physical Pain and the Goal of Aversively Stimulated Aggression." *Journal of Personality and Social Psychology* 40 (4) (April): 687–700.

Berkowitz, Leonard, and Eddie Harmon-Jones. 2004. "Toward an Understanding of the Determinants of Anger." *Emotion (Washington, D.C.)* 4 (2) (June): 107–30. doi:10.1037/1528-3542.4.2.107.

Berman, Mitchell. 2011. "Two Types of Retributivism." In *The Philosophical Foundations of Criminal Law*, edited by Stuart Duff, R. A, Green. Oxford University Press.

Blair, R J R. 2012. "Considering Anger from a Cognitive Neuroscience." *Cognitive Science* 3 (February): 65–74. doi:10.1002/wcs.154.

Blanchard, D. Caroline, and Robert J. Blanchard. 1984. "Affect and Aggression: An Animal Model Applied to Human Behavior." In *Advances in the Study of Aggression*, edited by Robert J Blanchard and D Caroline Blanchard, 1:1–62.

———. 1988. "Ethoexperimental Approaches to the Biology of Emotion." *Annual Review of Psychology*.

———. 2003a. “What Can Animal Aggression Research Tell Us about Human Aggression?” *Hormones and Behavior* 44 (3) (September): 171–177. doi:10.1016/S0018-506X(03)00133-8.

Blanchard, D.Caroline, and Robert J Blanchard. 2003b. “What Can Animal Aggression Research Tell Us about Human Aggression?” *Hormones and Behavior* 44 (3) (September): 171–177. doi:10.1016/S0018-506X(03)00133-8.

Blanchard, Robert J., D. Caroline Blanchard, T. Takahashi, and M. J. Kelley. 1977. “Attack and Defensive Behaviour in the Albino Rat.” *Animal Behaviour* 25: 622–634.

Boesch, Christophe. 1991. “The Effects of Leopard Predation on Grouping Patterns in Forest Chimpanzees.” *Behaviour* 117 (3): 220–242.

Boice, R. 1977. “Burrows of Wild and Albino Rats: Effects of Domestication, Outdoor Raising, Age, Experience, and Maternal State.” *Journal of Comparative and Physiological Psychology* 91 (3) (June): 649–61.

Bowles, S., and H. Gintis. 2004. “The Evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations.” *Theoretical Population Biology* 65 (1): 17–28.

———. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton Univ Pr.

Bowles, Samuel, Jung-Kyoo Choi, and Astrid Hopfensitz. 2003. “The Co-Evolution of Individual Behaviors and Social Institutions.” *Journal of Theoretical Biology* 223 (2) (July): 135–147. doi:10.1016/S0022-5193(03)00060-2.

Boyd, R. 1991. “Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds.” *Philosophical Studies*.

Boyd, Robert, Herbert Gintis, and Samuel Bowles. 2010. “Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare.” *Science (New York, N.Y.)* 328 (5978) (April 30): 617–20. doi:10.1126/science.1183665.

Bridewell, Will B., and Edward C. Chang. 1997. “Distinguishing between Anxiety, Depression, and Hostility: Relations to Anger-In, Anger-Out, and Anger Control.” *Personality and Individual Differences* 22 (4) (April): 587–590. doi:10.1016/S0191-8869(96)00224-3.

Brigandt, Ingo. 2009. “Natural Kinds in Evolution and Systematics: Metaphysical and Epistemological Considerations.” *Acta Biotheoretica* 57 (1-2) (July): 77–97. doi:10.1007/s10441-008-9056-7.

Brosnan, Sarah F., and Frans B M de Waal. 2003. “Monkeys Reject Unequal Pay.” *Nature* 425: 297–299.

Brown, Rachael L. 2013. “Identifying Behavioral Novelty.” *Biological Theory* (December 5). doi:10.1007/s13752-013-0150-y.

Buckholtz, Joshua W, Christopher L Asplund, Paul E Dux, David H Zald, John C Gore, Owen D Jones, and René Marois. 2008. “The Neural Correlates of Third-Party Punishment.” *Neuron* 60 (5) (December 10): 930–40. doi:10.1016/j.neuron.2008.10.016.

Burge, Tyler. 2010. *Origins of Objectivity*. OUP Oxford.

Calder, Andrew J, Jill Keane, Andrew D Lawrence, and Facundo Manes. 2004. “Impaired Recognition of Anger Following Damage to the Ventral Striatum.” *Brain□: A Journal of Neurology* 127 (Pt 9) (September): 1958–69. doi:10.1093/brain/awh214.

Canteras, Newton S. 2002. “The Medial Hypothalamic Defensive System: Hodological Organization and Functional Implications.” *Pharmacology, Biochemistry, and Behavior* 71 (3) (March): 481–91.

Carlsmith, Kevin M, John M Darley, and Paul H Robinson. 2002a. “Why Do We Punish□? Deterrence and Just Deserts as Motives for Punishment.” *Journal of Personality and Social Psychology* 83 (2): 284–299. doi:10.1037//0022-3514.83.2.284.

Carlsmith, Kevin M. 2006. “The Roles of Retribution and Utility in Determining Punishment.” *Journal of Experimental Social Psychology* 42 (4) (July): 437–451. doi:10.1016/j.jesp.2005.06.007.

———. 2008. “On Justifying Punishment: The Discrepancy Between Words and Actions.” *Social Justice Research* 21 (2) (May 3): 119–137. doi:10.1007/s11211-008-0068-x.

Carlsmith, Kevin M., John M. Darley, and Paul H. Robinson. 2002b. “Why Do We Punish?: Deterrence and Just Deserts as Motives for Punishment.” *Journal of Personality and Social Psychology* 83 (2): 284–299. doi:10.1037//0022-3514.83.2.284.

Carlsmith, KM, and JM Darley. 2008. “Psychological Aspects of Retributive Justice.” *Advances in Experimental Social Psychology*.

Carrier, DR, and MH Morgan. 2014. “Protective Buttressing of the Hominin Face.” *Biological Reviews*.

Carruthers, P. 2004. “On Being Simple Minded.” *American Philosophical Quarterly*.

Carver, Charles S, and Eddie Harmon-jones. 2009. “Anger Is an Approach-Related Affect□: Evidence and Implications.” *Psychological Bulletin* 135 (2): 183–204. doi:10.1037/a0013965.

Cheney, DL, and RM Seyfarth. 1989. “Redirected Aggression and Reconciliation among Vervet Monkeys, Cercopithecus Aethiops.” *Behaviour* 110 (1): 258–275.

Chevalier-Skolnikoff, S. 1974. “The Ontogeny of Communication in the Stumptail Macaque (Macaca Arctoides).”

Chevalier-Skolnikoff, S. 1973. “Facial Expression of Emotion in Nonhuman Primates.” In *… and Facial Expression: A Century of …*, 11–89.

Clark, Jason. 2009. “Relations of Homology between Higher Cognitive Emotions and Basic Emotions.” *Biology & Philosophy* 25 (1) (May 23): 75–94. doi:10.1007/s10539-009-9170-1.

Clore, GL. 1994. “Why Emotions Are Felt.” In *The Nature of Emotion: Fundamental Questions*.

Clore, GL, and A Ortony. 1993. “Where Does Anger Dwell.” *Perspectives on Anger ….*

Côté, Sylvana M, Tracy Vaillancourt, John C LeBlanc, Daniel S Nagin, and Richard E Tremblay. 2006. “The Development of Physical Aggression from Toddlerhood to Pre-Adolescence: A Nation Wide Longitudinal Study of Canadian Children.” *Journal of Abnormal Child Psychology* 34 (1) (February): 71–85. doi:10.1007/s10802-005-9001-z.

Critchfield, T S, and S H Kollins. 2001. “Temporal Discounting: Basic Research and the Analysis of Socially Important Behavior.” *Journal of Applied Behavior Analysis* 34 (1) (January): 101–22. doi:10.1901/jaba.2001.34-101.

Crockett, Molly J. 2013. “Models of Morality.” *Trends in Cognitive Sciences* 17 (8) (August): 363–6. doi:10.1016/j.tics.2013.06.005.

Cushman, Fiery. 2013. “Action, Outcome, and Value: A Dual-System Framework for Morality.” *Personality and Social Psychology Review* 17 (3) (August): 273–92. doi:10.1177/1088868313495594.

Cushman, Fiery, Kurt Gray, Allison Gaffey, and Wendy Berry Mendes. 2012. “Simulating Murder: The Aversion to Harmful Action.” *Emotion* 12 (1) (February): 2–7. doi:10.1037/a0025071.

Cushman, Fiery, Liane Young, and Joshua D Greene. 2007. “Our Multi-System Moral Psychology□: Towards a Consensus View”: 1–20.

Daly, Martin, and Margo Wilson. 1988. *Homicide*. Transaction Publishers.

Damasio, A.R. 1994. *Descartes’ Error*. New York: Putnam.

Darley, J M, K M Carlsmith, and P H Robinson. 2000. “Incapacitation and Just Deserts as Motives for Punishment.” *Law and Human Behavior* 24 (6) (December): 659–83.

Darwin, C. 1872. *The Expression of Emotions in Animals and Man*. London: J. Murray Publ.

Davidson, D. 1963. “Actions, Reasons, and Causes.” *The Journal of Philosophy*.

———. 1986. “Rational Animals. Actions and Events: Perspectives on the Philosophy of Donald Davidson.”

Dawkins, R. 2006. *The Selfish Gene*.

Defensor, Erwin B, Michael J Corley, Robert J Blanchard, and D Caroline Blanchard. 2012. “Facial Expressions of Mice in Aggressive and Fearful Contexts.” *Physiology & Behavior* 107 (5) (December 5): 680–5. doi:10.1016/j.physbeh.2012.03.024.

Delgado, Jose M. R. 1967. “Aggression and Defense Under Cerebral Radio Control.” In *Aggression and Defense: Neural Mechanisms and Social Patterns*, edited by Carmine D. Clemente and Donald B. Lindsley, 171–193. University of California Press.

———. 1968. “Offensive-Defensive Behaviour in Free Monkeys and Chimpanzees Induced by Radio Stimulation of the Brain.” In *Aggressive Behavior*, edited by S. Garattini and E. B. Sigg, 109–119. New York: John Wiley and Sons.

Dennett, Daniel C. 1984. “Cognitive Wheels: The Frame Problem Minds, Machines and Evolution.” In *Minds, Machines and Evolution*, edited by Hookway.

Denson, TF. 2009. “Angry Rumination and the Self-Regulation of Aggression.” In *Psychology of Self-Regulation: Cognitive, Affective, and Motivational Processes*, edited by Joseph P. Forgas, Roy F. Baumeister, and Dianne M. Tice, 233–248. New York: Taylor & Francis.

DeScioli, Peter, and Robert Kurzban. 2009. “Mysteries of Morality.” *Cognition* 112 (2) (August): 281–99. doi:10.1016/j.cognition.2009.05.008.

DeWall, C. Nathan, Roy F. Baumeister, Tyler F. Stillman, and Matthew T. Gailliot. 2007. “Violence Restrained: Effects of Self-Regulation and Its Depletion on Aggression.” *Journal of Experimental Social Psychology* 43 (1) (January): 62–76. doi:10.1016/j.jesp.2005.12.005.

Dretske, Fred I. 1991. *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

———. 1999. *Knowledge and the Flow of Information*. Center for the Study of Language and Information Publications (CSLI).

Duff, A. 2001. *Punishment, Communication, and Community*.

Eibl-Eibesfeldt, I. 1961. “The Fighting Behavior of Animals.” *Scientific American* 205: 112–122.

———. 1973. “The Expressive Behavior of the Deaf-and-Blind-Born.” *Social Communication and Movement*: 163–194.

———. 1979. “Human Ethology: Concepts and Implications for the Sciences of Man.” *Behavioral and Brain Sciences* 2 (01): 1–26.

Ekman, Paul. 1977. “Biological and Cultural Contributions to Body and Facial Movement.” In *Anthropology of the Body*, edited by John Blacking, 34–84.

———. 1999. “Basic Emotions.” In *The Handbook of Cognition and Emotion*, edited by Tim Dalgleish and Mick Power, 45–60. Sussex, UK: John Wiley & Sons, Ltd.

———. 2003. *Emotion Revealed: Understanding Faces and Feelings*. Phoenix Press.

Ekman, Paul, W. V Friesen, M. O’Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, et al. 1987. “Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion.” *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology* 53 (4): 712.

Ekman, Paul, W. V. Friesen, and R. C. Simons. 1985. “Is the Startle Reaction an Emotion?” *Journal of Personality and Social Psychology* 49 (5): 1416.

Ekman, Paul, and WV Friesen. 1971. “Constants across Cultures in the Face and Emotion.” *Journal of Personality and Social ….*

Ekman, Paul, R. W Levenson, and W. V Friesen. 1983. “Autonomic Nervous System Activity Distinguishes among Emotions.” *Science* 221 (4616): 1208–1210.

Ekman, Paul, and others. 1971. *Universals and Cultural Differences in Facial Expressions of Emotion*. University of Nebraska Press Lincoln.

Ekman, Paul, E R Sorenson, and W V Friesen. 1969. “Pan-Cultural Elements in Facial Displays of Emotion.” *Science*.

Elster, John. 1990. “Norms of Revenge.” *Ethics* 100 (4): 862–885.

Elster, Jon. 1999. *Strong Feelings: Emotion, Addiction, and Human Behavior*. MIT Press.

Ereshefsky, Marc. 2007. “Psychological Categories as Homologies: Lessons from Ethology.” *Biology & Philosophy* 22 (5) (September 29): 659–674. doi:10.1007/s10539-007-9091-9.

———. 2012. “Homology Thinking.” *Biology & Philosophy* 27 (3) (March 10): 381–400. doi:10.1007/s10539-012-9313-7.

Evans, Jonathan St.B.T. 2003. “In Two Minds: Dual-Process Accounts of Reasoning.” *Trends in Cognitive Sciences* 7 (10) (October): 454–459. doi:10.1016/j.tics.2003.08.012.

Ewer, RF. 1971. “The Biology and Behaviour of a Free-Living Population of Black Rats (Rattus Rattus).”

Fabiansson, Emma C, and Thomas F Denson. 2012. “The Effects of Intrapersonal Anger and Its Regulation in Economic Bargaining.” *PloS One* 7 (12) (January): e51595. doi:10.1371/journal.pone.0051595.

Fehr, Ernst, and Urs Fischbacher. 2004. “Third-Party Punishment and Social Norms.” *Evolution and Human Behavior* 25 (2) (March): 63–87. doi:10.1016/S1090-5138(04)00005-4.

Fehr, Ernst, and Simon Gächter. 2002. “Altruistic Punishment in Humans.” *Nature* 415 (6868) (January 10): 137–40. doi:10.1038/415137a.

Feinberg, J. 1970. “Doing & Deserving; Essays in the Theory of Responsibility.”

Felsten, Gary. 1996. “Pergamon HOSTILITY , STRESS AND SYMPTOMS OF DEPRESSION” 21 (4): 461–467.

Ferreira, A, L G Dahlöf, and S Hansen. 1987. “Olfactory Mechanisms in the Control of Maternal Aggression, Appetite, and Fearfulness: Effects of Lesions to Olfactory Receptors, Mediodorsal Thalamic Nucleus, and Insular Prefrontal Cortex.” *Behavioral Neuroscience* 101 (5) (October): 709–17, 746.

Flannelly, K J, and D H Thor. 1978. “Territorial Aggression of the Rat to Males Castrated at Various Ages.” *Physiology & Behavior* 20 (6) (June): 785–9.

Flannelly, Kevin, and Richard Lore. 1975. “Dominance-Subordinance in Cohabiting Pairs of Adult Rats□: Effects on Aggressive Behavior” 1: 331–340.

Fokkema, D S, J M Koolhaas, and J van der Gugten. 1995. “Individual Characteristics of Behavior, Blood Pressure, and Adrenal Hormones in Colony Rats.” *Physiology & Behavior* 57 (5) (May): 857–62.

Foot, P. 1978. “The Problem of Abortion and the Doctrine of Double Effect.” In *Virtues and Vices*. Oxford.

Fraassen, Bas C. Van. 1980. *The Scientific Image*. Clarendon Press.

Frank, R.H. 1988. *Passions within Reason: The Strategic Role of the Emotions*. New York: Norton.

Frankfurt, Harry G, and G Frankfurt. 2014. “North American Philosophical Publications The Problem of Action X . THE PROBLEM OF ACTION” 15 (2): 157–162.

Fridlund, AJ. 1991. “Sociality of Solitary Smiling: Potentiation by an Implicit Audience.” *Journal of Personality and Social Psychology*.

Frijda, Nico H. 1986. *The Emotions*. *The Emotions*. Vol. 1. Studies in Emotion and Social Interaction. Cambridge University Press. doi:10.1093/0199253048.001.0001.

———. 2010. “Impulsive Action and Motivation.” *Biological Psychology* 84 (3) (July): 570–9. doi:10.1016/j.biopsycho.2010.01.005.

Frijda, Nico H., Peter Kuipers, and Elisabeth ter Schure. 1989. “Relations among Emotion, Appraisal, and Emotional Action Readiness.” *Journal of Personality and Social Psychology* 57 (2): 212–228. doi:10.1037//0022-3514.57.2.212.

Gambaro, S, and AI Rabin. 1969. “Diastolic Blood Pressure Responses Following Direct and Displaced Aggression after Anger Arousal in High-and Low-Guilt Subjects.” *Journal of Personality and Social Psychology*.

Gauker, Christopher. 2003. *Words Without Meaning*. MIT Press.

Geen, R G, D Stonner, and G L Shope. 1975. “The Facilitation of Aggression by Aggression: Evidence against the Catharsis Hypothesis.” *Journal of Personality and Social Psychology* 31 (4) (April): 721–6.

Geist, V. 1974. “On the Relationship of Social Evolution and Ecology in Ungulates.” *Integrative and Comparative Biology* 14 (1): 205–220. doi:10.1093/icb/14.1.205.

Gerra, G, A Zaimovic, and P Avanzini. 1997. “Neurotransmitter-Neuroendocrine Responses to Experimentally Induced Aggression in Humans: Influence of Personality Variable.” *Psychiatry* ….

Gibbard, A. 1992. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Harvard University Press.

Ginet, C. 1990. *On Action*.

Gläscher, Jan, Nathaniel Daw, Peter Dayan, and John P O’Doherty. 2010. “States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning.” *Neuron* 66 (4) (May 27): 585–95. doi:10.1016/j.neuron.2010.04.016.

Goldberg, Julie H., Jennifer S. Lerner, and Philip E. Tetlock. 1999. “Rage and Reason: The Psychology of the Intuitive Prosecutor.” *European Journal of Social Psychology* 29: 781–795.

Goodall, J. 1986. “The Chimpanzees of Gombe: Patterns of Behavior.”

Grant, C, and M R A Chance. 1957. “Rank Order in Caged Rats.”

Grant, EC, and MRA Chance. 1958. “Rank Order in Caged Rats.” *Animal Behaviour*.

Greene, Joshua. 2003. “From Neural ‘Is’ to Moral ‘Ought’: What Are the Moral Implications of Neuroscientific Moral Psychology?” *Nature Reviews. Neuroscience* 4 (10) (October): 846–9. doi:10.1038/nrn1224.

Greene, Joshua D. 2008. “The Secret Joke of Kant’s Soul.” In *Moral Psychology, Vol. 3, The Neuroscience of Morality: Emotion, Disease, and Development*, edited by Walter Sinnott-Armstrong, 35–80. Cambridge: MIT Press.

Greene, Joshua D, Fiery a Cushman, Lisa E Stewart, Kelly Lowenberg, Leigh E Nystrom, and Jonathan D Cohen. 2009. “Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment.” *Cognition* 111 (3) (June): 364–71. doi:10.1016/j.cognition.2009.02.001.

Griffiths, Paul E, and Andrea Scarantino. 2004. “Emotions in the Wild□: The Situated Perspective on Emotion *”: 1–28.

Griffiths, Paul E. 1997. *What Emotions Really Are: The Problem of Psychological Categories*. Vol. 1997. University of Chicago Press.

———. 2001. “Genetic Information: A Metaphor in Search of a Theory.” *Philosophy of Science*: 394–412.

———. 2007. “The Phenomena of Homology.” *Biology & Philosophy* 22 (5) (October 10): 643–658. doi:10.1007/s10539-007-9090-x.

———. 2010. “Basic Emotions, Complex Emotions, Machiavellian Emotions.” *Royal Institute of Philosophy Supplement* 52 (January 8): 39–67. doi:10.1017/S1358246100007888.

Griffiths, Paul E., and Edouard Machery. 2008. “Innateness, Canalization, and ‘Biologicizing the Mind.’” *Philosophical Psychology* 21 (3) (June): 397–414. doi:10.1080/09515080802201146.

Griffiths, PE. 2006. “Function, Homology, and Character Individuation*.” *Philosophy of Science* 73 (1): 1–25.

Hamilton, W D. 1964. “The Genetical Evolution of Social Behaviour. I.” *Journal of Theoretical Biology* 7 (1) (July): 1–16.

Hampton, J. 1992. “An Expressive Theory of Punishment.” In *Retributivism and Its Critics*.

Harmon-Jones, E., C. K. Peterson, and C. Harmon-Jones. 2010. “Anger, Motivation, and Asymmetrical Frontal Cortical Activations.” *International Handbook of Anger*: 61–78.

Hart, HLA. 1967. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford University Press.

Hess, WR. 1954. “Diencephalon, Autonomic and Extrapyramidal Functions.”

Hitchcock, E, and V Cairns. 1973. “Amygdalotomy.” *Postgraduate Medical Journal* 49 (578) (December): 894–904.

Hokanson, J E, M Burgess, and M F Cohen. 1963. “Effects of Displaced Aggression on Systolic Blood Pressure.” *Journal of Abnormal Psychology* 67 (3) (September): 214–8.

Hokanson, J E, and S Shetler. 1961. “The Effect of Overt Aggression on Physiological Arousal Level.” *Journal of Abnormal and Social Psychology* 63 (2) (September): 446–8.

Hokanson, JE, and Michael Burgess. 1962. “The Effects of Status, Type of Frustration, and Aggression on Vascular Processes.” *The Journal of Abnormal and Social Psychology* 65 (4): 232–237.

Hubbard, J. A., L. J. Romano, M. D. McAuliffe, and M. T. Morrow. 2010. “Anger and the Reactive–Proactive Aggression Distinction in Childhood and Adolescence.” In *International Handbook of Anger*, edited by Michael Potegal, Gerhard Stemmler, and Charles D Spielberger, 231–239. Springer New York.

Hurley, Susan, Greg Currie, Martin Davies, Julia Driver, Peter Godfrey-smith, Mark Greenberg, Celia Heyes, et al. 2003. “Animal Action in the Space of Reasons” 18 (3): 231–256.

Izard, C E. 1994. “Innate and Universal Facial Expressions: Evidence from Developmental and Cross-Cultural Research.” *Psychological Bulletin* 115 (2) (March): 288–99.

Izard, C. E. 1971. “The Face of Emotion.”

Izard, Carroll E., Elizabeth a. Hembree, and Robin R. Huebner. 1987. “Infants’ Emotion Expressions to Acute Pain: Developmental Change and Stability of Individual Differences.” *Developmental Psychology* 23 (1): 105–113. doi:10.1037//0012-1649.23.1.105.

Kahane, Guy. 2011. “Evolutionary Debunking Arguments.” *Nous* 45 (1) (March): 103–125. doi:10.1111/j.1468-0068.2010.00770.x.

———. 2012. “On the Wrong Track: Process and Content in Moral Psychology.” *Mind & Language* 27 (5) (November): 519–545. doi:10.1111/mila.12001.

Kahane, Guy, Katja Wiech, Nicholas Shackel, Miguel Farias, Julian Savulescu, and Irene Tracey. 2012. “The Neural Basis of Intuitive and Counterintuitive Moral Judgment.” *Social Cognitive and Affective Neuroscience* 7 (4) (April): 393–402. doi:10.1093/scan/nsr005.

Kamm, F.M. 1993. *Morality, Mortality*. Vol. 2. Oxford.

Kelly, D. 2011. *Yuck!: The Nature and Moral Significance of Disgust*. Cambridge: MIT Press.

Kolunie, J M, and J M Stern. 1995. “Maternal Aggression in Rats: Effects of Olfactory Bulbectomy, ZnSO4-Induced Anosmia, and Vomeronasal Organ Removal.” *Hormones and Behavior* 29 (4) (December): 492–518. doi:10.1006/hbeh.1995.1285.

Kroon, FW. 1985. “Theoretical Terms and the Causal View of Reference.” *Australasian Journal of Philosophy* (February 2014): 37–41.

Kruk, M R. 1991. “Ethology and Pharmacology of Hypothalamic Aggression in the Rat.” *Neuroscience and Biobehavioral Reviews* 15 (4) (January): 527–38.

Kruk, M R, C E Van der Laan, J Mos, a M Van der Poel, W Meelis, and B Olivier. 1984. "Comparison of Aggressive Behaviour Induced by Electrical Stimulation in the Hypothalamus of Male and Female Rats." *Progress in Brain Research* 61 (January): 303–14. doi:10.1016/S0079-6123(08)64443-X.

Kruk, M R, and a M van der Poel. 1980. "Is There Evidence for a Neural Correlate of an Aggressive Behavioural System in the Hypothalamus of the Rat?" *Progress in Brain Research* 53 (January): 385–90. doi:10.1016/S0079-6123(08)60077-1.

Kruk, M R, a M Van der Poel, W Meelis, J Hermans, P G Mostert, J Mos, and a H Lohman. 1983. "Discriminant Analysis of the Localization of Aggression-Inducing Electrode Placements in the Hypothalamus of Male Rats." *Brain Research* 260 (1) (January 31): 61–79.

Lazarus, R. S. 1991. *Emotion and Adaptation*. Oxford University Press, USA.

LeDoux, Joseph. 1998. *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Vol. 25. Simon and Schuster.

Lerner, J. S, J. H Goldberg, and P. E Tetlock. 1998. "Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility." *Personality and Social Psychology Bulletin* 24 (6): 563–574.

Lerner, Jennifer S, and Larissa Z Tiedens. 2006. "Portrait of The Angry Decision Maker□: How Appraisal Tendencies Shape Anger ' S Influence on Cognition." *Journal of Behavioral Decision Making* 137: 115–137. doi:10.1002/bdm.515.

Levenson, R. W. 1992. "Autonomic Nervous System Differences among Emotions." *Psychological Science* 3 (1): 23.

Levenson, R. W, Paul Ekman, and W. V Friesen. 1990. "Voluntary Facial Action Generates Emotion-Specific Autonomic Nervous System Activity." *Psychophysiology* 27 (4): 363–384.

Leyhausen, Paul. 1979. *Cat Behavior: The Predatory and Social Behavior of Domestic and Wild Cats*. Translated by Batrbara A. Tonkin. 1St Editio. Taylor & Francis / Garland STPM Press.

Lipp, H. P., and R. W. Hunsperger. 1978. "Threat, Attack and Flight Elicited by Electrical Stimulation of the Ventromedial Hypothalamus of the Marmoset Monkey." *Brain, Behavior and Evolution* 15: 260–293.

Looney, TA, and PS Cohen. 1982. "Aggression Induced by Intermittent Positive Reinforcement." *Neuroscience & Biobehavioral Reviews*.

Lorenz, K. 2003. *The Foundations of Ethology*.

Lorenz, Konrad. 1957. "The Nature of Instinct." In *Instinctive Behavior: The Development of a Modern Concept*. New York: International Universities Press.

Machamer, P., L. Darden, and C. F Craver. 2000. “Thinking about Mechanisms.” *Philosophy of Science* 67 (1): 1–25.

Machery, Edouard, and Ron Mallon. 2010. “Evolution of Morality.” In *The Moral Psychology Handbook*, edited by John M. Doris. Oxford University Press.

Mallon, R., and S. P. Stich. 2000. “The Odd Couple: The Compatibility of Social Construction and Evolutionary Psychology.” *Philosophy of Science*: 133–154.

Mallon, Ron, and JM Weinberg. 2006. “Innateness as Closed Process Invariance.” *Philosophy of Science* 73 (3): 323–344.

Marler, P, A Dufty, and R Pickert. 1986. “Vocal Communication in the Domestic Chicken: II. Is a Sender Sensitive to the Presence and Nature of a Receiver?” *Animal Behaviour*.

Mason, Kelby. 2011. “Moral Psychology And Moral Intuition: A Pox On All Your Houses.” *Australasian Journal of Philosophy* 89 (3) (September): 441–458. doi:10.1080/00048402.2010.506515.

Mauss, Iris B., Crystal L. Cook, and James J. Gross. 2007. “Automatic Emotion Regulation during Anger Provocation.” *Journal of Experimental Social Psychology* 43 (5) (September): 698–711. doi:10.1016/j.jesp.2006.07.003.

Maynard Smith, John. 1974. “The Theory of Games and the Evolution of Animal Conflicts.” *Journal of Theoretical Biology*.

Maynard Smith, John, and D. Harper. 2003. *Animal Signals*. Oxford University Press, USA.

Mayr, E. 1974. “Behavior Programs and Evolutionary Strategies.” *American Scientist* 62 (6): 650–659.

McCullough, Michael E, Robert Kurzban, and Benjamin a Tabak. 2012. “Cognitive Systems for Revenge and Forgiveness.” *The Behavioral and Brain Sciences* (December 5): 1–15. doi:10.1017/S0140525X11002160.

Mele, AR. 2000. “Goal-Directed Action: Teleological Explanations, Causal Theories, and Deviance.” *Noûs* 34 (May): 279–300.

Milburn, MA, SD Conrad, F Sala, and S Carberry. 1995. “Childhood Punishment, Denial, and Political Attitudes.” *Political Psychology*.

Milburn, Michael A., Miho Niwa, and Marcus D. Patterson. 2014. “Authoritarianism, Anger, and Hostile Attribution Bias: A Test of Affect Displacement.” *Political Psychology* 35 (2) (August 19): 225–243. doi:10.1111/pops.12061.

Millikan, RG. 1993. “What Is Behavior?” In *White Queen Psychology and Other Essays for Alice*.

Moore, MS. 2010. *Placing Blame: A Theory of the Criminal Law*.

Morris, H. 1968. “Persons and Punishment.” *The Monist*.

Moyer, KE. 1976. “The Psychobiology of Aggression.”

Nakao, Hisashi, and Edouard Machery. 2012. “The Evolution of Punishment.” *Biology & Philosophy* 27 (6) (September 2): 833–850. doi:10.1007/s10539-012-9341-3.

Nelissen, Rob M A, and Marcel Zeelenberg. 2009. “Moral Emotions as Determinants of Third-Party Punishment□: Anger , Guilt , and the Functions of Altruistic Sanctions” 4 (7): 543–553.

Newman, M G, and a a Stone. 1996. “Does Humor Moderate the Effects of Experimentally-Induced Stress?” *Annals of Behavioral Medicine□: A Publication of the Society of Behavioral Medicine* 18 (2) (June): 101–9. doi:10.1007/BF02909582.

Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford University Press, USA.

Nola, R. 1980. “Fixing the Reference of Theoretical Terms.” *Philosophy of Science*.

Nozick, R. 1974. *Anarchy, State, and Utopia*. Basic Books.

Nozick, Robert. 1981. *Philosophical Explanations*. Cambridge, MA: Belknap Press of Harvard University Press.

O’Neill, Elizabeth. “Innateness as the Insensitivity of the Appearance of a Trait with Respect to Specified Environmental Variation.”

Öhman, A. 1986. “Face the Beast and Fear the Face: Animal and Social Fears as Prototypes for Evolutionary Analyses of Emotion.” *Psychophysiology*.

Olsen, RW. 1969. “Agonistic Behavior of the Short-Tailed Shrew (Blarina Brevicauda).” *Journal of Mammalogy* 50 (3): 494–500.

Owen, R. 1846. *Lectures on the Comparative Anatomy and Physiology of the Vertebrate Animals*. Vol. 2. Printed for Longman, Brown, Green, and Longmans.

Pacherie, Elisabeth. 2006. “Towards a Dynamic Theory of Intentions”: 145–167.

Panksepp, Jaak. 1971. “Aggression Elicited by Electrical Stimulation of the Hypothalamus in Albino Rats.” *Physiology & Behavior* 6 (4) (April): 321–9. doi:10.1016/0031-9384(71)90163-6.

———. 1994. “The Basics of Basic Emotion.” In *The Nature of Emotion: Fundamental Questions*.

———. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. 1st ed. Oxford University Press, USA.

Panksepp, Jaak, and Lucy Biven. 2012. *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*. W. W. Norton & Company.

Panksepp, Jaak, and MR Margaret R Zellner. 2004. "Towards A Neurobiologically Based Unified Theory of Aggression." *REVUE INTERNATIONALE DE …* 17 (2): 37–61.

Parfit, Derek. 2011. *On What Matters, Vol. 2*. Edited by Oxford University Press.

Parr, LA, BM Waller, SJ Vick, and KA Bard. 2007a. "Classifying Chimpanzee Facial Expressions Using Muscle Action." *Emotion* 7 (1): 172–181. doi:10.1037/1528-3542.7.1.172.Classifying.

———. 2007b. "Classifying Chimpanzee Facial Expressions Using Muscle Action." *Emotion (Washington, DC)* 7 (1): 172–181. doi:10.1037/1528-3542.7.1.172.Classifying.

Parr, Lisa a., Mirit Cohen, and Frans De Waal. 2005. *Influence of Social Context on the Use of Blended and Graded Facial Displays in Chimpanzees*. *International Journal of Primatology*. Vol. 26. doi:10.1007/s10764-005-0724-z.

Paxton, JM, Tommaso Bruni, and JD Greene. 2013. "Are 'counter-Intuitive' Deontological Judgments Really Counter-Intuitive?: An Empirical Reply to Kahane et al.(2012)." *Social Cognitive and …*: 1–9.

Pedersen, William C, Thomas F Denson, R Justin Goss, Eduardo a Vasquez, Nicholas J Kelley, and Norman Miller. 2011. "The Impact of Rumination on Aggressive Thoughts, Feelings, Arousal, and Behaviour." *The British Journal of Social Psychology / the British Psychological Society* 50 (Pt 2) (June): 281–301. doi:10.1348/014466610X515696.

Petersen, Michael Bang, Aaron Sell, John Tooby, and Leda Cosmides. 2010. "And Criminal Justice□: A Recalibrational Theory of Punishment and Reconciliation" (2).

Pillutla, MM, and JK Murnighan. 1996. "Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers." *Organizational Behavior and Human Decision …*.

Pojman, LP. 2005. "A Defense of the Death Penalty." *Contemporary Debates in Applied Ethics*.

Pollock, John. 1986. *Contemporary Theories of Knowledge*. Totowa, NJ: Rowman & Littlefield.

Pollock, John L. 1987. "Defeasible Reasoning." *Cognitive Science* 11 (4) (October 11): 481–518. doi:10.1207/s15516709cog1104_4.

Potegal, M. 1979. “The Reinforcing Value of Several Types of Aggressive Behavior: A Review.” *Aggressive Behavior* 5 (4): 353–373.

———. 1994. “Aggressive Arousal!” *The Dynamics of Aggression: Biological and Social Processes in Dyads and Groups*: 73.

Potegal, M., and G. Stemmler. 2010. “Constructing a Neurology of Anger.” *International Handbook of Anger*: 39–59.

Potegal, Michael. 1992. “Time Course of Aggressive Arousal in Female Hamsters and Male Rats.” *Behavioral and Neural Biology* 58 (2) (September): 120–4.

Potegal, Michael, and Liliana TenBrink. 1984. “Behavior of Attack-Primed and Attack-Satiated Female Golden Hamsters (Mesocricetus Auratus).” *Journal of Comparative Psychology* 98 (1): 66–75. doi:10.1037//0735-7036.98.1.66.

Preuss, T.M. 2007. *Evolution of Nervous Systems*. *Null*. Vol. null. Elsevier. doi:10.1016/B0-12-370878-8/00005-7.

Prinz, Jesse. 2007. *The Emotional Construction of Morals*. Vol. 22. Oxford University Press.

Prinz, Jesse J. 2004. *Gut Reactions□: A Perceptual Theory of Emotion: A Perceptual Theory of Emotion*. Oxford University Press.

Putnam, Hilary. 1969. “Brains and Behaviour.” In *Analysis*, edited by Ned Block, 30:1–19. Mind, Language and Reality. Blackwell.

Quinn, W. 1985. “The Right to Threaten and the Right to Punish.” *Philosophy & Public Affairs*.

Railton, P. 2012. “That Obscure Object, Desire.”

Rawls, John. 1955. “Two Concepts of Rules.” *The Philosophical Review* 64 (1) (January): 3. doi:10.2307/2182230.

Remane, Adolph. 1971. “Die Grundlagen Des Natürlichen Systems: Der Vergleichenden Anatomie Und Der Phylogenetik.”

Richerson, P. J, and R. Boyd. 2004. *Not by Genes Alone: How Culture Transformed Human Evolution*. University of Chicago Press.

Rieppel, Olivier. 2005. “Modules, Kinds, and Homology.” *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 304 (1) (January 15): 18–27. doi:10.1002/jez.b.21025.

Roberts, W. W., M. L. Steinberg, and L. W. Means. 1967. “Hypothalamic Mechanisms for Sexual, Aggressive and Other Motivational Behaviors in the Oppossum Didelphis Virginiana.” *Journal of Comparative Physiology and Psychology* 64: 1–15.

Roseman, Ira J. 2004. “Appraisals, rather than Unpleasantness or Muscle Movements, Are the Primary Determinants of Specific Emotions.” *Emotion (Washington, D.C.)* 4 (2) (July): 145–50; discussion 151–5. doi:10.1037/1528-3542.4.2.145.

Rozin, Paul, and April E. Fallon. 1987. “A Perspective on Disgust.” *Psychological Review* 94 (1): 23–41.

Rozin, Paul, Laura Lowery, and Jonathan Haidt. 1999. “The CAD Triad Hypothesis□: A Mapping Between Three Moral Emotions (Contempt , Anger , Disgust) and Three Moral Codes (Community , Autonomy , Divinity).” *Journal of Personality and Social Psychology* 76 (4): 574–586.

Rubenstein, D I. 1981. “ROLE ASSESSMENT , RESERVE STRATEGY , AND ACQUISITION OF INFORMATION IN ASYMMETRIC ANIMAL CONFLICTS over a Resource of Equal Value to Each Individual” (ii): 221–240.

Russell, J a, and B Fehr. 1994. “Fuzzy Concepts in a Fuzzy Hierarchy: Varieties of Anger.” *Journal of Personality and Social Psychology* 67 (2) (August): 186–205.

Sanfey, A. G, J. K Rilling, J. A Aronson, L. E Nystrom, and J. D Cohen. 2003. “The Neural Basis of Economic Decision-Making in the Ultimatum Game.” *Science* 300 (5626): 1755–1758.

Scheffler, Samuel. 1994. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Clarendon Press.

Scherer, K. 2003. “Vocal Communication of Emotion: A Review of Research Paradigms.” *Speech Communication* 40 (1-2) (April): 227–256. doi:10.1016/S0167-6393(02)00084-5.

Schroeder, T. 2004. “Three Faces of Desire.”

Scott, John. 1976. *Aggression*. University of Chicago press.

Sehon, SR. 1994. “Teleology and the Nature of Mental States.” *American Philosophical Quarterly* 31 (1): 63–72.

———. 2007. “Goal-Directed Action and Teleological Explanation.” In *Causation and Explanation*, 155–170.

Sell, Aaron. “Applying Adaptationism to Human Anger: The Recalibrational Theory Aaron Sell Center for Evolutionary Psychology University of California, Santa Barbara.”

Sell, Aaron N. 2011. “The Recalibrational Theory and Violent Anger.” *Aggression and Violent Behavior* 16 (5) (September): 381–389. doi:10.1016/j.avb.2011.04.013.

Sell, Aaron, John Tooby, and Leda Cosmides. 2009. “Formidability and the Logic of Human Anger.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (35) (September 1): 15073–8. doi:10.1073/pnas.0904312106.

Shaikh, M B, a Steinberg, and a Siegel. 1993. “Evidence That Substance P Is Utilized in Medial Amygdaloid Facilitation of Defensive Rage Behavior in the Cat.” *Brain Research* 625 (2) (October 22): 283–94.

Siegel, Allan. 2004. *Neurobiology of Aggression and Rage*. 1st ed. Informa Healthcare.

Siegel, Allan, S Bhatt, Rekha Bhatt, and SS Zalcman. 2010. “Limbic, Hypothalamic and Periaqueductal Gray Circuitry and Mechanisms Controlling Rage and Vocalization in the Cat.” *Handbook of Behavioral …* 19: 243–253. doi:10.1016/B978-0-12-374593-4.00024-3.

Siegel, Allan, T a Roeling, T R Gregg, and M R Kruk. 1999. “Neuropharmacology of Brain-Stimulation-Evoked Aggression.” *Neuroscience and Biobehavioral Reviews* 23 (3) (January): 359–89.

Siegel, Allan, and Jeff Victoroff. 2009. “Understanding Human Aggression: New Insights from Neuroscience.” *International Journal of Law and Psychiatry* 32 (4): 209–15. doi:10.1016/j.ijlp.2009.06.001.

Singer, Peter. 2005. “Ethics and Intuitions.” *The Journal of Ethics* 9 (3): 331–352.

Sinnott-Armstrong, Walter. 2003. “Consequentialism” (May 20).

———. 2008. “Framing Moral Intuitions.” In *Moral Psychology, Vol. 2, The Cognitive Science of Morality: Intuition and Diversity*, edited by Walter Sinnott-armstrong. Cambridge: MIT Press.

Smith, JM, and GR Price. 1973. “Lhe Logic of Animal Conflict.” *Nature*.

Smits, Dirk J.M., and Peter Kuppens. 2005. “The Relations between Anger, Coping with Anger, and Aggression, and the BIS/BAS System.” *Personality and Individual Differences* 39 (4) (September): 783–793. doi:10.1016/j.paid.2005.02.023.

Sober, Elliott. 2009. *Philosophy Of Biology*. Second. Westview Press.

Sosa, Ernest. 2007. “Experimental Philosophy and Philosophical Intuition.” *Philosophical Studies* 132 (1): 99–107. doi:10.1007/sll098-006-9050-3.

Srivastava, Joydeep, and Francine Espinoza. 2009. “Coupling and Decoupling of Unfairness and Anger in Ultimatum Bargaining” 489 (December 2008): 475–489. doi:10.1002/bdm.

Stanford, PK, and P Kitcher. 2000. “Refining the Causal Theory of Reference for Natural Kind Terms.” *Philosophical Studies* 97 (1): 99–129.

Stanovich, Keith E. 2004. *The Robot’s Rebellion*. University of Chicago press.

Sterelny, K. 1999. “Situated Agency and the Descent of Desire.” *Where Biology Meets Psychology: Philosophical ….*

———. 2001. *The Evolution of Agency and Other Essays*.

Street, Sharon. 2006. “A Darwinian Dilemma for Realist Theories of Value.” *Philosophical Studies* 127 (1): 109–166. doi:10.1007/sl.

Strobel, Alexander, Jan Zimmermann, Anja Schmitz, Martin Reuter, Stefanie Lis, Sabine Windmann, and Peter Kirsch. 2011. “Beyond Revenge: Neural and Genetic Bases of Altruistic Punishment.” *NeuroImage* 54 (1) (January 1): 671–80. doi:10.1016/j.neuroimage.2010.07.051.

Tamir, Maya, Christopher Mitchell, and James J Gross. 2008. “Hedonic and Instrumental Motives in Anger Regulation.” *Psychological Science* 19 (4) (April): 324–8. doi:10.1111/j.1467-9280.2008.02088.x.

Tappolet, Christine. 2013. “Emotion , Motivation , and Action□: The Case of Fear.” In *Oxford Handbook of Emotion*. Oxford University Press. doi:10.1093/oxfordhb/9780199235018.003.0015.

Tedeschi, James T. 1994. *Violence, Aggression, and Coercive Actions*. American Psychological Association.

Tinbergen, N. 1963. “On Aims and Methods of Ethology.” *Zeitschrift Für Tierpsychologie*.

Tooby, John, and Leda Cosmides. 1990. “The Past Explains the Present.” *Ethology and Sociobiology* 11 (4-5) (July): 375–424. doi:10.1016/0162-3095(90)90017-Z.

Trivers, RL. 1971. “The Evolution of Reciprocal Altruism.” *Quarterly Review of Biology*.

Tyson, Paul D. 1998. “Physiological Arousal, Reactive Aggression, and the Induction of an Incompatible Relaxation Response.” *Aggression and Violent Behavior* 3 (2) (June): 143–158. doi:10.1016/S1359-1789(97)00002-5.

Vaillant, John. 2010. *The Tiger: A True Story of Vengeance and Survival*. Knopf Doubleday Publishing Group.

Van ’t Wout, Mascha, René S Kahn, Alan G Sanfey, and André Aleman. 2006. “Affective State and Decision-Making in the Ultimatum Game.” *Experimental Brain Research* 169 (4) (March): 564–8. doi:10.1007/s00221-006-0346-5.

Vavova, Katia. 2014. “Debunking Evolutionary Debunking.” In *Oxford Studies in Metaethics*, 1–37. Oxford University Press.

Viger, C. 2000. “Where Do Dennett’s Stances Stand? Explaining Our Kind of Mind.” *Dennett’s Philosophy: A Comprehensive Assessment*.

Vitiello, B, and D M Stoff. 1997. “Subtypes of Aggression and Their Relevance to Child Psychiatry.” *Journal of the American Academy of Child and Adolescent Psychiatry* 36 (3) (March): 307–15. doi:10.1097/00004583-199703000-00008.

Walen, Alec. 2014. “Retributive Justice” (June 18).

Walletschek, H, and a Raab. 1982. “Spontaneous Activity of Dorsal Raphe Neurons during Defensive and Offensive Encounters in the Tree-Shrew.” *Physiology & Behavior* 28 (4) (April): 697–705.

Wasman, M, and JP Flynn. 1962. “Directed Attack Elicited from Hypothalamus.” *Archives of Neurology*.

Weinshenker, Naomi J, and Allan Siegel. 2002. “Bimodal Classification of Aggression: Affective Defense and Predatory Attack.” *Aggression and Violent Behavior* 7 (3) (May): 237–250. doi:10.1016/S1359-1789(01)00042-8.

Wenzel, JW. 1992. “Behavioral Homology and Phylogeny.” *Annual Review of Ecology and Systematics* 23 (1992): 361–381.

Wingfield, J C, S Lynn, and K K Soma. 2001. “Avoiding the ‘Costs’ of Testosterone: Ecological Bases of Hormone-Behavior Interactions.” *Brain, Behavior and Evolution* 57 (5) (May): 239–51.

Wingfield, JC, RE Hegner, AM Dufty Jr, and GF Ball. 1990. “The‘ Challenge Hypothesis’: Theoretical Implications for Patterns of Testosterone Secretion, Mating Systems, and Breeding Strategies.” *American Naturalist*.

Woodfield, Andrew. 1976. *Teleology*. Cambridge University Press.

Woodworth, C H. 1971. “Attack Elicited in Rats by Electrical Stimulation of the Lateral Hypothalamus.” *Physiology & Behavior* 6 (4) (April): 345–53.

Zaibert, L. 2006. “Punishment and Revenge.” *Law and Philosophy* 25 (1): 81–118.

Zillmann, Dolf. 1979. *Hostility and Aggression*. L. Erlbaum Associates.